Anno LVI
Numero 2
Agosto 2007

Elettronica e telecomunicazioni

Rai Centro Ricerche e Innovazione Tecnologica

Rai Eri

Editoriale

Ravello - Qualità Tecnica della Musica

Tecniche di visualizzazione stereoscopica basate sulla frequenza:
anaglifo e Infitec™

Olofonia - una ripresa sonora
di tutto ciò che ci circonda

Alta Definizione: display 1080p

La tecnologia 3D-HDTV basata su DLP®

Algoritmi e tecnologie per il riconoscimento vocale:
stato dell'arte e sviluppi futuri

Edizione ottimizzata per la stampa.
La rivista è disponibile su web
alla URL www.crit.rai.it/eletel.htm

Anno LVI
N° 2
Agosto 2007

Rivista quadrimestrale a cura della Rai

Direttore responsabile
Gianfranco Barbieri

Comitato direttivo
Gino Alberico
Marzio Barbero
Mario Cominetti
Alberto Morello
Mario Stroppiana

Redazione
Marzio Barbero
Gemma Bonino

Indice

Editoriale

ing. Gianfranco **Barbieri**
Direttore di
"Elettronica e
Telecomunicazioni"

La stragrande maggioranza degli schermi televisivi esistenti oggi sul mercato viene pubblicizzata come prodotto avente i requisiti per la ricezione della TV ad Alta Definizione. Come, tuttavia, è stato illustrato in una serie di articoli pubblicati in passato su questa rivista la denominazione TV ad Alta definizione (HDTV) non è di per sé sufficiente ad individuare in modo univoco uno Standard; esiste infatti una famiglia di Standard ciascuno dei quali è caratterizzato dal numero di righe, dalla frequenza di ripetizione di quadro, dalla struttura interallacciata o progressiva della trama. Spesso l'acquirente di un nuovo televisore si trova in difficoltà di fronte alla varie proposte dei centri di vendita: HD Ready o Full HD, 50 Hz o 100Hz, LCD o Plasma. A complicare la problematica si aggiunge l'eventualità di una varietà di "processamenti" digitali interni al display (tipica è la conversione dello standard di scansione negli schermi HD Ready che si rende necessaria per rendere compatibile la visione di programmi trasmessi in Full HD). L'articolo "Alta Definizione: display 1080p" intende fornire al lettore alcune essenziali informazioni che lo aiutino a districarsi nella scelta di uno schermo dell'ultima generazione.

Con l'avvento della TV ad Alta definizione ed il consolidamento del formato panoramico 16/9 nasce l'esigenza di fornire un'immagine sonora spaziale che dia all'utente un effetto di presenza all'interno della scena e che sia all'altezza della qualità video. Il Centro Ricerche della Rai sta sperimentando nuovi sistemi di ripresa sonora che vadano incontro alle suddette esigenze pur rispettando gli obiettivi di ottimizzazione dei costi di produzione e di agevole interfacciamento con le infrastrutture esistenti. L'articolo:"Olofonia: una ripresa sonora di tutto ciò che ci circonda" illustra il sistema olofonico e riporta i primi risultati ottenuti.

Un secondo filone di articoli comparsi recentemente nella nostra rivista ha riguardato la TV Stereoscopica. Uno dei principali problemi della stereoscopia consiste nella difficoltà di veicolare all'occhio corretto il rispettivo canale in fase di visualizzazione. Nel caso di stereoscopia basata su una coppia di immagini fisse o in movimento, si dà per scontato il tipo di ripresa (con una coppia di telecamere affiancate) e si assume che in qualche modo si possa alimentare l'apparato di visualizzazione con la coppia di segnali stereoscopici eventualmente prelevandola da un canale trasmissivo. Le tecniche di visualizzazione sono, invece, numerose e il dibattito è ancora aperto tra gli addetti ai lavori poiché nessuna si impone in assoluto per le sue qualità. Nell'articolo:"Tecniche di visualizzazione stereoscopica basate sulla frequenza: anàglifo e Infitec" si descrivono due delle suddette tecniche, entrambe basate sulla frequenza della luce

Ravello

Qualità Tecnica della Musica

Dal 15 al 17 giugno si è tenuto alla Villa Rufolo
di Ravello, la Città della Musica,
il 2° incontro dedicato alla Qualità Tecnica della Musica
organizzato dalla Direzione Strategie Tecnologiche della Rai.

All'incontro, realizzato con il patrocinio della Provincia di Salerno,
hanno partecipato docenti universitari,
operatori di comunicazione, musicisti ed
esperti di tecnologie in rappresentanza della ricerca e dell'industria.

Fra i contributi presentati dal Centro Ricerche Rai al convegno, che si è
tenuto il 16 giugno, uno riguardava la ripresa olofonica, oggetto di uno
degli articoli in questo numero.

Da sinistra: Antonio Bottiglieri (dirigente Rai)
Luigi Rocchi (direttore di Strategie Tecnologiche Rai),
Paolo Imperato (sindaco di Ravello) e
Luigi Nicolais (ministro per le Riforme e l'Innovazione nella Pubblica Amministrazione)

Tecniche di visualizzazione stereoscopica basate sulla frequenza: **anàglifo e Infitec**

ing. Mario **Muratori**

Rai
Centro Ricerche e
InnovazioneTecnologica
Torino

Sommario
Uno dei primi metodi introdotti per la visualizzazione stereoscopica con sovrapposizione delle immagini fu l'anàglifo.
La tecnica si adatta bene alle tecnologie di stampa a colori e fu largamente adottata durante il boom della fotografia della fine del 1800.
L'anàglifo è stato utilizzato in seguito anche alla cinematografia e vi sono proposte recenti anche su DVD; di conseguenza vi furono tentativi di applicarlo anche alla televisione. In questo caso, però, mostra evidenti limiti che ne hanno fortemente limitato ogni prospettiva, confinandolo ad iniziative promozionali o sperimentali.
La tecnica denominata col marchio registrato di Infitec sfrutta anch'essa, in modo molto ingegnoso, la frequenza della luce, ottenendo un sistema di elevatissime prestazioni proposto soprattutto per presentazioni visive di qualità.

1. Introduzione

Uno dei principali problemi della stereoscopia consiste nella difficoltà di veicolare all'occhio corretto il rispettivo canale in fase di visualizzazione.

Nel caso di stereoscopia "tradizionale"[Nota 1] si dà per scontato il tipo di ripresa - con una coppia di telecamere affiancate - e si assume che in qualche modo si possa alimentare l'apparato di visualizzazione con la coppia di segnali stereoscopici[Nota 2] - eventualmente prelevandola da un canale trasmissivo.

Nota 1 - Qui si indica con l'aggettivo "tradizionale" la stereoscopia basata su una coppia di immagini fisse o in movimento, in contrapposizione ad altre forme più evolute di stereoscopia, in particolare quella basata sull'integral imaging.
Nota 2 - Oppure, in alcuni sistemi, con un segnale video e una mappa di profondità che, in linea di principio, si possono ricavare relativamente facilmente dalla coppia di segnali stereoscopici.

Al contrario, le tecniche di visualizzazione sono numerose, ognuna caratterizzata da fattori positivi e fattori negativi. Nessuna si impone in assoluto per le sue qualità, alcune si adattano meglio di altre alla configurazione di visione[Nota 3], altre sono invece decisamente poco utilizzate[Nota 4].

Normalmente si visualizzano i due canali sovrapposti spazialmente [1], quindi il problema consiste nel separare i due segnali prima che giungano all'occhio dell'utente.

Alcuni metodi si basano su tecniche "sistemistiche", per esempio multiplando nel tempo i due segnali e separandoli con occhiali *shutter*, le cui lenti diventano trasparenti od opache in sincronia con il segnale visualizzato. Altre tecniche si basano sulle proprietà della luce, in particolare la polarizzazione o la frequenza[Nota 5].

In questo articolo si descrivono due tecniche basate sulla frequenza della luce: il metodo dell'*anàglifo* e la tecnica denominata *Infitec*™.

2. Anàglifo

2.1. Generalità

Il termine deriva dal tardo latino *anaglyphus*, a sua volta derivato dal greco ανα–γλυφοσ, cesellato.

Secondo il Perucca, la dizione corretta è: *metodo degli anàglifi*. Si tratta di un metodo per l'osservazione di immagini stereoscopiche basato sull'impiego di *colori complementari* [2]: le due immagini costituenti la coppia stereoscopica sono colorate con colori complementari, ad esempio rosso e ciano.

Osservandole contemporaneamente con occhiali aventi come lenti dei filtri colorati[Nota 6] - corrispondenti ai colori usati per le immagini, nell'esempio sopra: rosso e ciano - ogni occhio vede solo l'immagine colorata con lo stesso colore del filtro postogli di fronte, mentre il colore complementare viene annullato (visto come nero) poiché assorbito dal filtro. In questo modo si riesce a far pervenire a ciascun occhio l'immagine corrispondente, separando le due componenti stereoscopiche.

La visione stereoscopica si ottiene perché la funzionalità della fusione sensoriale, operata dal cervello e fondamentale per il riconoscimento degli oggetti, si basa principalmente sul riconoscimento delle forme, che vengono riconosciute anche se hanno colori falsati.

La percezione del rilievo invece deriva dalla stereopsi, ossia dall'interpretazione delle diffe-

Nota 3 - Le tecniche a proiezione sono preferibili in caso di multiutenza in grandi sale, p.es. cinema. Alcuni sistemi autostereoscopici non permettono la visione multiutente, quindi sono più adatti alla visualizzazione di dati che all'uso televisivo.

Nota 4 - Le tecniche oggigiorno più diffuse sono basate sulla polarizzazione della luce e sulla multiplazione nel tempo con occhiali attivi (shutter). Gli HDM (Head Mount Display) sono utilizzati solo in applicazioni di nicchia soprattutto per il costo e la modalità di uso monoutente. I display autostereoscopici incontrano ancora difficoltà tecniche di realizzazione [7]. I sistemi basati su anàglifo e tecnica Infitec sono trattati nel seguito dell'articolo.

Nota 5 - Si noti che l'ampiezza dell'onda elettromagnetica costituente la luce supporta l'informazione di luminosità del pixel, mentre l'occhio è insensibile alla fase della luce incidente. Pertanto questi due parametri non possono essere utilizzati per discriminare i due canali della coppia stereoscopica.

Nota 6 - Un filtro ottico è uno strumento che trasmette selettivamente la luce con particolari proprietà (una particolare lunghezza d'onda, una gamma di colore o di luce), bloccando invece le rimanenti. Sono comunemente usati in fotografia e in molti strumenti ottici. I filtri ad assorbimento sono di solito fabbricati in vetro a cui sono stati aggiunti vari materiali inorganici o organici. Questi componenti assorbono alcune lunghezze d'onda della luce lasciandone passare altre. A volte si utilizzano materiali plastici (per esempio policarbonato o acrilico) per produrre filtri di gel, più leggeri e meno costosi di quelli vitrei.

renze tra le due immagini dovute alle differenti viste prospettiche, che sono visibili grazie alla separazione tra i canali operata dai filtri di cui sono equipaggiati gli occhiali.

Si noti che assieme alla fusione sensoriale si verifica anche una "fusione colorimetrica": i colori percepiti dai due occhi si sommano ottenendo una riproduzione del colore per sintesi additiva. Per questo motivo è possibile utilizzare coppie formate da immagini mancanti di colori primari (il verde ed il blu nel canale colorato di rosso, il rosso nell'altro), ottendo come risultato complessivo una resa colorimetrica relativamente accettabile, ancorché obiettivamente non ottimale.

2.2. La produzione di immagini anàglife

Il metodo classico per produrre un anàglifo consiste nel riprendere la scena con una coppia di mezzi di ripresa appaiati, posti ad una certa distanza tra loro, filtrare con filtri opportunamente colorati le due immagini, sovrapponendole in fase di stampa o di visualizzazione in modo da evidenziare le disparità orizzontali derivanti dal parallasse di ripresa.

In linea di principio, il filtraggio colorimetrico può essere effettuato in qualsiasi fase del processo: dalla ripresa alla fase di visualizzazione o stampa; in genere, quando le immagini vengono elaborate al calcolatore con appositi applicativi, per esempio per realizzare ritocchi, filtri, effetti, composizioni e così via, risulta conveniente effettuare il filtraggio in questa fase, in quanto gli applicativi più evoluti hanno gli strumenti necessari per ottenere anàglifi (layer, regolazione dei livelli dei primari, ecc.).

Anticamente si utilizzavano solamente le coppie di primari rosso-verde o rosso-blu. In questo modo la riproduzione del colore era di scarsa qualità, perché con due soli primari non si può riprodurre una gamma di colori sufficiente.

Attualmente si preferisce invece adottare la coppia di colori complementari rosso-ciano; la complementarietà di questi due colori permette una migliore reiezione di quello indesiderato; inoltre il ciano è composto dai due primari verde e blu, quindi complessivamente si utilizzano tutti i tre primari ottenendo una migliore resa colorimetrica, in particolare sull'incarnato.

In genere l'immagine sinistra è filtrata per rimuovere il blu e il verde: rimane il rosso che attraversa il filtro rosso (che negli occhiali per anaglifo è posto davanti all'occhio sinistro). L'immagine destra è filtrata per rimuovere il rosso: rimangono le componenti blu e verdi che attraversano il filtro ciano.

2.3. La visualizzazione

Per la visualizzazione degli anàglifi si utilizzano occhiali dotati di filtri colorati; in figura 1 si riporta una tipica caratteristica in frequenza per il filtro rosso e quello blu.

Davanti all'occhio sinistro si trova un filtro rosso, davanti a quello destro sono stati utilizzati filtri di diverso colore a seconda degli orientamenti del periodo: nel passato si sono adottati filtri verdi oppure blu, oggigiorno è preferito un filtro di colore ciano capace di far passare le componenti sia blu che verdi (figura 2).

Gli occhiali più semplici, equipaggiati con semplici filtri di gelatina, non compensano la differenza

Fig. 1 – Caratteristiche spettrali di filtri colorati blu e rosso (fonte: [3]).

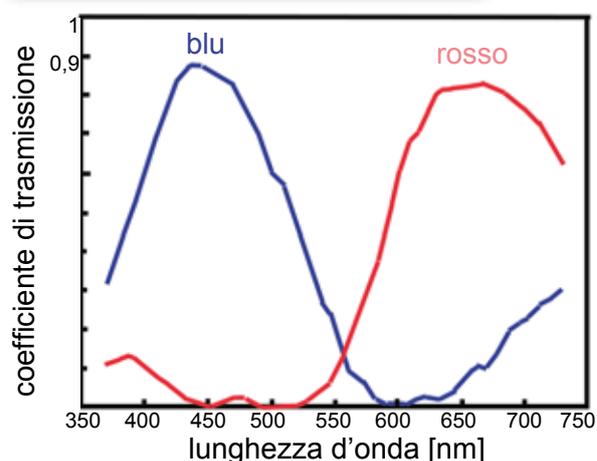



Fig. 2 – Occhiali per per la visualizzazione degli anàglifi.

Fig. 4 – Occhiali di tipo anachrome.
(fonte: sito en.wikipedia.org/wiki/Anaglyph_image)

di lunghezza focale, pari a 250 nm, tra il rosso e il ciano. Nei sistemi ottici tale differenza provoca la distorsione chiamata aberrazione cromatica[Nota 6] (figura 3).

In stereoscopia l'effetto è che l'immagine sinistra, colorata di rosso, appare sfocata rispetto a quella destra – che normalmente è dominante relativamente alla messa a fuoco.

Tecniche migliorative

I filtri di maggiore spessore e di materiale opportuno[Nota 7] possono essere prodotti in modo da fungere da lenti con una potenza ottica di valore tale da compensare l'aberrazione cromatica.

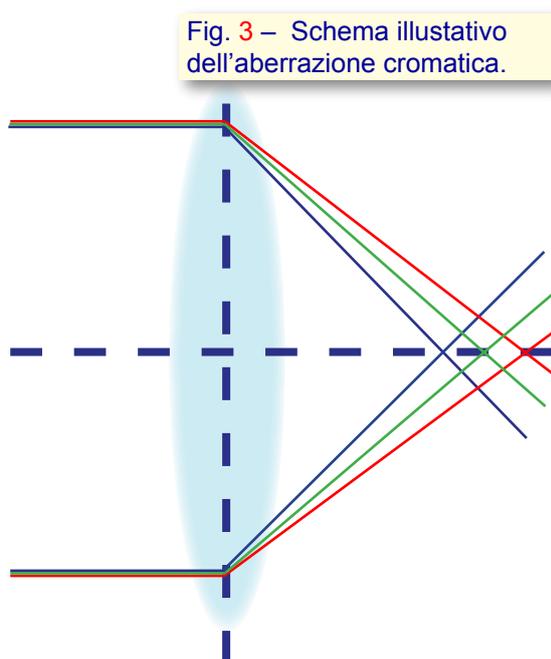



Fig. 3 – Schema illustrativo dell'aberrazione cromatica.

Normalmente si adotta una diottria pari a +1/2 sulla lente rossa; in tal modo però si ingrandisce leggermente l'immagine sinistra, introducendo un'altra distorsione potenzialmente fastidiosa.

Un altro tipo di ottimizzazione consiste nell'adottare un filtro ciano che permetta anche il passaggio di una piccola percentuale di rosso per migliorare la resa cromatica dell'incarnato; ovviamente il filtraggio effettuato sulle immagini ne deve tenere conto.

In una variante denominata anachrome, oltre ad adottare il filtro ciano modificato (figura 4),

Nota 6 - In ottica, l'aberrazione cromatica è un difetto nella formazione dell'immagine dovuta al diverso valore di rifrazione delle diverse lunghezze d'onda che compongono la luce che passa attraverso il mezzo ottico. In pratica succede che per radiazioni policromatiche le componenti con lunghezza d'onda più corta (blu) vengono rifratte maggiormente rispetto a quelle con lunghezza d'onda più lunga creando una dispersione dell'immagine sul piano focale. Questo effetto è particolarmente fastidioso nei microscopi e nei telescopi. L'aberrazione cromatica viene corretta, ma non del tutto eliminata, utilizzando un sistema ottico composto da lenti multiple di materiali a diversa dispersione in modo che le differenze tra gli angoli di rifrazione per la stessa lunghezza d'onda si annullino tra loro (sistemi ottici acromatici).

Nota 7 - Vi sono diversi materiali plastici trasparenti e con proprietà ottiche notevoli; si segnalano, come esempio, il polimetilmetacrilato (Plexiglas), infrangibile e più trasparente del vetro, e il policarbonato, usato nelle lenti infrangibili per occhiali.

in fase di stampa viene limitata anche l'entità delle disparità orizzontali in modo che il bordo degli oggetti appaia solo leggermente sdoppiato, ottenendo delle immagini quasi compatibili 2D (appaiono come piccoli difetti di allineamento di stampa), ma riducendo in proporzione l'effetto prospettico.

In [3] è descritto un metodo matematico di ottimizzazione volto ad ottenere l'immagine anàglifa migliore possibile dal punto di vista percettivo, operando sia sul filtraggio colorimetrico che sulla composizione delle due componenti filtrate.

Il metodo tiene conto delle caratteristiche spettrali dei filtri degli occhiali e dell'emissione delle sorgenti luminose costituenti i pixel del display, nonché delle curve di sensibilità dell'occhio medio in funzione della frequenza.

A detta dell'Autore, tale processo di ottimizzazione dà risultati migliori rispetto ai metodi empirici descritti in precedenza. Tuttavia sembra piuttosto specifico per il display considerato e quindi non appare direttamente utilizzabile in campo televisivo dove non si possono imporre limiti troppo stringenti sul display da utilizzare.

2.4. Evoluzione e applicazioni non televisive

Gli inizi

Nel 1853, W. Rollman per primo illustrò la tecnica dell'anàglifo usando disegni colorati in blu e rosso (su fondo nero) e degli occhiali con lenti degli stessi colori.

Nel 1858 Joseph D'Almeida incominciò a proiettare immagini con la lanterna magica, usando filtri rossi e verdi.

Louis Ducas du Hauron[Nota 8] fu il primo, nel 1891, ad ottenere anàglifi stampati. Il suo metodo consisteva nello stampare sullo stesso foglio due negativi, uno colorato di blu o verde e l'altro di rosso.

Fig. 5 – Immagine stereoscopica generata col metodo degli anàglifi; il canale sinistro (per l'occhio sinistro) è colorato in rosso, quello destro in ciano.

La produzione iconografica recente

Recentemente, immagini stereoscopiche in anàglifo hanno avuto una certa diffusione in Internet, su CD e in stampa, risultando relativamente gradevoli alla vista.

Ciò anche grazie ai filtri adottati (rosso - ciano), che permettono di utilizzare tutti e tre i primari colorimetrici e quindi offrono una migliore riproduzione del colore, soprattutto delle tonalità della pelle (incarnato), rispetto alle coppie di filtri rosso-blu e rosso-verde usati in passato.

Nota 8 - Louis Ducas Du Hauron (1837-1920) fu uno scienziato francese che diede importanti contributi allo sviluppo della fotografia a colori. Nel suo libro "Les Couleurs en Photographie" (1869) propose il metodo sottrattivo per la fotografia a colori. Al tempo le sue teorie non vennero messe in pratica per la mancanza di materiali adatti, ma sono alla base delle moderne tecniche fotografiche. Nel 1891 brevettò il metodo degli anàglifi per la fotografia stereoscopica. Fu premiato nel 1900 con la "Progress Medal" dalla Royal Photographic Society per il suo lavoro nel campo della fotografia a colori.

Molto interesse, anche per la novità del contenuto, hanno suscitato le immagini di Marte diffuse dalla NASA (figura 6).

La cinematografia stereoscopica a metà del XX secolo

Nel periodo 1952-1954 i film stereoscopici ebbero una certa fortuna con più di trenta titoli distribuiti in tutto il mondo, anche se successivamente furono soppiantati dal sistema Cinemascope a grande schermo.

Tuttavia, secondo la documentazione reperita, la quasi totalità di tali film era proiettata con la tecnica della proiezione con luce polarizzata e non in anàglifo.

L'interesse per i film stereoscopici risorse più volte durante gli anni '60, '70 e '80 grazie a semplificazioni nella tecnica di proiezione, in particolare con i sistemi SpaceVision, usato da Oboler[Nota 9], e Stereovision utilizzato da Silliphant[Nota 10] che utilizzavano un solo proiettore. In ambedue i casi si trattava ancora, però, di sistemi basati sulla polarizzazione della luce.

Quest'ultimo rimase il formato principale per la cinematografia stereoscopica fino all'avvento dell'IMAX, nella cui versione stereoscopica (IMAX-3D), peraltro, non si adotta la tecnica anàglifa[Nota 11].

La cinematografia recente

Nel 2003, il film Spy Kids 3D basato sulla tecnica degli anàglifi ebbe un buon successo commerciale nei cinema. Al contrario, nel 2005, il film The Adventures of SharkBoy & LavaGirl, realizzato anch'esso in anàglifo dagli stessi creatori di Spy Kids, non riuscì a coprire gli investimenti.

Per confronto, nello stesso periodo i film basati su tecnica polarizzata incontrarono maggior successo di pubblico e commerciale; si citano ad esempio Chicken Little della Disney del 2005 e diversi film in formato IMAX tra cui Polar Express del 2004.

Nota 9 - Arch Oboler (1909-1987), famoso per le sue opere radiofoniche, nel 1960 produsse, scrisse e diresse il film Fantastic Invasion of Plantet Earth 3-D, per la proiezione del quale utilizzò la tecnica chiamata SpaceVision che richiedeva un solo proiettore (anziché una coppia) ed era basata sulla polarizzazione della luce. Nel 1952 scrisse, produsse e diresse il film Bwana Devil, considerato il primo film stereoscopico a colori prodotto negli Stati Uniti, proiettato con la tecnica chiamata Natural Vision, basata su due pellicole.

Nota 10 - Allan Silliphant scrisse, produsse e diresse il film The Stewardesses (1969, 1971; www.thestewardesses.com) che fu il film tridimensionale ad ottenere il maggior guadagno della storia (forse anche perché conteneva contenuti sessuali rilevanti per il tempo, essendo classificato come "softcore"). Il film fu girato in pellicola 35mm e proiettato con un formato speciale chiamato Stereovision sviluppato dalla Stereovision International, azienda dello stesso Silliphant. La visualizzazione avveniva con tecnica a luce polarizzata e non fu mai proiettato in anàglifo.

Nota 11 - Nei cinema IMAX-3D si utilizzano sistemi a polarizzazione e a multiplazione temporale.

Fig. 6 – Immagine della superficie di Marte in anàglifo.
(fonte: sito mars.jpl.nasa.gov/MPF/mpf/anaglyph-arc.html)

Da quanto emerge dalla documentazione reperita, sembrerebbe che si possano sintetizzare le vicende del mercato cinematografico sostenendo che la tecnica con luce polarizzata, migliore dal punto di vista tecnico, è preferita nei cinema, dove la maggior complessità sistemistica e logistica (uso di proiettori speciali o accoppiati, gestione degli occhiali equipaggiati con filtri polarizzatori) non rappresenta un grosso problema, mentre il metodo dell'anàglifo è preferito per l'home cinema, dove l'esigenza di ridurre i costi e l'assenza di apparati specifici rappresentano forti limitazioni.

2.5. Anàglifo e televisione

La tecnica anaglifica, anche se non è compatibile con la televisione bidimensionale "normale" per via dello sdoppiamento delle immagini colorate, sembrerebbe relativamente adatta all'uso televisivo poiché richiede un solo canale trasmissivo: viene quindi evitata l'esigenza di maneggiare due canali (eventualmente sincronizzati), cosa necessaria invece in altre tecniche stereoscopiche, nonché il dimezzamento della frequenza di quadro e le difficoltà nel ricavare la sincronizzazione per gli occhiali attivi richiesti dalla tecnica field-sequential.

Tuttavia le trasmissioni sperimentali finora effettuate a più riprese da diversi operatori non hanno avuto una buona accoglienza presso il pubblico.

Ciò sembrerebbe dovuto principalmente al fatto che in tutti i sistemi televisivi utilizzati per la trasmissione e la registrazione, DVD compresi, i segnali di crominanza vengono filtrati, e in certi sistemi piuttosto notevolmente, riducendone la definizione. Ricostruendo l'immagine a partire da tali segnali la definizione complessiva non è elevata e questo può degradarne la qualità soggettiva.

La scarsa fedeltà nella riproduzione del colore e la non elevata saturazione riducono ulteriormente la qualità soggettiva al punto che tale tipo di visualizzazione non gode di elevato gradimento da parte dell'utenza.

La sperimentazione svolta presso il Centro Ricerche

Presso il Centro Ricerche Rai è sorta l'esigenza di mostrare l'effetto reale delle problematiche sopra esposte.

Pertanto si è sviluppato un programma software per la generazione di una sequenza anàglifa a partire da una coppia stereoscopica[Nota 12]. L'algoritmo prevede:

- una trasformazione lineare per rappresentare i segnali nel dominio dei primari colorimetrici RGB
- il filtraggio colorimetrico
- la combinazione delle due sequenze filtrate in un'unica sequenza anàglifa
- una nuova trasformazione lineare per rappresentare i segnali nel formato YC_rC_b (luminanza e crominanza) adottato per la visualizzazione.

Si noti che le ultime due operazioni si possono commutare.

Il filtraggio adottato è molto semplice e consiste nell'annullare il contributo dei primari non voluti nel canale preso in considerazione. Ciò corrisponde ad applicare una coppia di filtri - del tipo passa basso per il rosso e passa alto per il ciano - caratterizzati da una attenuazione nulla in banda passante, attenuazione infinita in banda attenuata e con la banda di transizione ricadente completamente nella banda compresa tra il verde (λ=550 nm circa) ed il rosso (λ=650 nm circa).

Tale tipo di filtraggio è senz'altro rozzo in confronto al metodo di ottimizzazione descritto in [3], ma si ritiene che tale ottimizzazione migliori la resa colorimetrica, mentre in campo televisivo i maggiori problemi nascono dal filtraggio delle

Nota 12 - I segnali televisivi sono disponibili in formato digitale, il che permette di effettuare agevolmente elaborazioni tramite calcolatore elettronico.

componenti crominanza e sono questi gli artefatti che ci si propone di evidenziare.

Analisi dell'algoritmo

Anche se il segnale televisivo a colori è generato sotto forma delle tre componenti relative ai tre primari colorimetrici (rosso R, verde G, blu B), normalmente i segnali televisivi sono memorizzati ed elaborati nel formato "luminanza più due segnali di crominanza", dove questi ultimi sono filtrati per limitarne la banda e sottocampionati.

Per questo lavoro si è adottato il formato ritenuto il massimo della qualità per le normali utilizzazioni[Nota 13], che è quello definito dalla raccomandazione ITU-R BT.601-5 [4] in cui si prevede che i segnali di crominanza siano sottocampionati di un fattore due rispetto al segnale di luminanza[Nota 14].

Nella stessa raccomandazione si riportano anche le matrici di conversione tra il dominio RGB e il formato YC_rC_b, che rappresentano semplici trasformazioni lineari interpretabili come un cambio di sistema di riferimento colorimetrico. In notazione matriciale, la trasformazione si può rappresentare nel modo seguente:

(1) $[Y\ C_r\ C_b]^T = [M]\ [R\ G\ B]^T$

Dove:

- $[Y\ C_r\ C_b]^T$ rappresenta il segnale video nelle tre componenti luminanza più crominanza.
- $[M]$ è la matrice di trasformazione tra dominio RGB e formato YC_rC_b riportata in [4], comprensiva delle necessarie normalizzazioni
- $[R\ G\ B]^T$ rappresenta il segnale video nelle tre componenti relative ai primari colorimetrici R, G e B.

La matrice [M] è invertibile: pertanto il segnale è trasformabile di nuovo esattamente nel formato RGB di partenza applicando la matrice inversa:

(2) $[R\ G\ B]^T = [M]^{-1}\ [Y\ C_r\ C_b]^T$

Dove $[M]^{-1}$ è la matrice inversa di [M].

Il concetto è facilmente comprensibile se si pensa che la moltiplicazione matriciale in questo caso corrisponde ad un cambio di sistema di riferimento colorimetrico: è ovvio che cambiando il riferimento il vettore rappresentato non si modifichi.

E' anche ovvio, tuttavia, che se si modifica il valore delle componenti in un certo dominio, non sarà possibile ricostruire i valori originali applicando la trasformazione inversa. In particolare si deve tenere in conto che i vettori che compaiono nelle relazioni indicate rappresentano dei segnali, il cui valore, funzione del tempo e dello spazio, viene modificato quando si applicano operazioni quali per esempio un filtraggio nel dominio delle frequenze.

Il segnale nel dominio dei primari RGB dato dalla (2) può essere considerato la somma della componente a bassa frequenza (apice BF) e della componente ad alta frequenza (apice AF) del segnale rappresentato in termini di componenti YC_rC_b:

(3a) $[R\ G\ B]^T =$

$$[M]^{-1}\ [Y\ C_r\ C_b]^T_{BF} + [M]^{-1}\ [Y\ C_r\ C_b]^T_{AF}$$

Se la componente di alta frequenza viene modificata, per esempio da un filtraggio nel dominio delle frequenze che annulli le componenti di crominanza in alta frequenza, è chiaro che non si potrà ricostruire esattamente il segnale RGB

Nota 13 - Per usi particolari è stato standardizzato anche il formato 4:4:4 dove le tre componenti non sono filtrate.
Nota 14 - Il formato è noto anche come 4:2:2. Si noti che si considerano qui segnali digitali, che permettono di applicare direttamente elaborazioni con algoritmi numerici.

di partenza:

(3b) $[R\ G\ B]^T \neq$

$[M]^{-1}\ [Y\ C_r\ C_b]^T_{BF} + [M]^{-1}\ [Y\ 0\ 0]^T_{AF}$

Tenendo conto di quanto detto, si analizzi l'algoritmo implementato per ottenere l'immagine anàglifa, che in notazione matriciale, si può rappresentare nel seguente modo:

(4) $V_A =$

$[M]\ [F]_S\ [M]^{-1}\ [V]_S + [M]\ [F]_D\ [M]^{-1}\ [V]_D =$

$[T]_S\ [V]_S + [T]_D\ [V]_D$

Dove:

- $V = [Y\ C_r\ C_b]^T$ rappresenta il segnale video nelle tre componenti luminanza più crominanza.
- [F] è la matrice che rappresenta il filtraggio colorimetrico effettuato dando un peso opportuno alle componenti del segnale video. Gli apici S e D indicano rispettivamente il filtro per il canale sinistro e destro.
- $[T] = [M]\ [F]\ [M]^{-1}$ è la matrice che tiene conto delle due trasformazioni e del filtraggio colorimetrico. Gli apici S e D indicano rispettivamente il canale sinistro e quello destro.

In altre parole: il segnale relativo al singolo canale della coppia stereoscopica ($[V]_{S,D}$) viene trasformato in formato RGB tramite moltiplicazione matriciale con $[M]^{-1}$; in questo dominio si effettua il filtraggio colorimetrico implementato con la moltiplicazione per l'opportuna matrice $[F]_{S,D}$; il risultato dei filtraggi sulle due componenti stereo viene trasformato di nuovo nel formato $Y\ C_r\ C_b$ per la sua utilizzazione con i comuni apparati e le due componenti - derivate dall'elaborazione dei canali sinistro e destro - sono sommate per generare l'immagine anàglifa. Si noti che sono tutte operazioni lineari e che le ultime due possono essere commutate.

Come accennato in precedenza, il segnale televisivo si può rappresentare come somma di una componente a bassa frequenza e di una ad alta frequenza (nella trattazione che segue non ha importanza la frequenza di taglio né l'esatto andamento dei filtri, purché si considerino solo operazioni lineari):

(5) $[V] = [V]_{BF} + [V]_{AF} =$

$[Y\ C_r\ C_b]^T_{BF} + [Y\ C_r\ C_b]^T_{AF}$

Sostituendo la notazione riportata in (5) nella (4) si ottiene:

(6) $V_A =$

$[T]_S\ ([Y\ C_r\ C_b]^T_{BF\ S} + [Y\ C_r\ C_b]^T_{AF\ S}) +$

$[T]_D\ ([Y\ C_r\ C_b]^T_{BF\ D} + [Y\ C_r\ C_b]^T_{AF\ D}) =$

$[T]_S\ [Y\ C_r\ C_b]^T_{BF\ S} + [T]_D\ [Y\ C_r\ C_b]^T_{BF\ D} +$

$[T]_S\ [Y\ C_r\ C_b]^T_{AF\ S} + [T]_D\ [Y\ C_r\ C_b]^T_{AF\ D} =$
$V_{A\ BF} + V_{A\ AF}$

Si noti che la componente in bassa frequenza dell'immagine anàglifa è ricostruita perfettamente:

(7a) $V_{A\ BF} =$

$[T]_S\ [Y\ C_r\ C_b]^T_{BF\ S} + [T]_D\ [Y\ C_r\ C_b]^T_{BF\ D}$

Mentre la componente in alta frequenza - composta sia dalla luminanza che dalla crominanza - non è ricostruita perfettamente in quanto, in ottemperanza alla raccomandazione citata, le componenti di alta frequenza delle crominanze sono azzerate, ottenendo:

(7b) $V_{A\ AF} = [T]_S\ [Y\ 0\ 0]^T_{AF\ S} + [T]_D\ [Y\ 0\ 0]^T_{AF\ D}$

Ricordando le relazioni (3a) e (3b) si noti che il filtraggio sulle crominanze non permette la ricostruzione perfetta del segnale nel dominio

RGB[Nota 15] su cui si opera il filtraggio colorimetrico. La successiva trasformazione [M] diffonde ulteriormente gli artefatti di cui sono affetti i segnali RGB su tutte e tre le componenti - Y, C_r, e C_b - che vengono visualizzate.

Da questa analisi risulta che l'immagine anàglifa (figura 7a) è meno definita rispetto all'originale (figura 7b) perché solo la componente in bassa frequenza viene generata correttamente, e contiene degli artefatti in quanto la componente in alta frequenza non viene ricostruita perfettamente.

Fig. 7a – Immagine anàglifa.

Fig. 7b – Immagine originale.

3. Infitec™ [Nota 16]

3.1. Premessa

Secondo il principio della tricromia, ogni colore visibile può essere riprodotto con un'opportuna miscela di tre colori primari.

Il metodo è ampiamente utilizzato nelle due forme chiamate sintesi additiva quando si sommano contributi luminosi generati da apposite sorgenti colorate, e sintesi sottrattiva quando si usano dei filtri per eliminare contributi luminosi di certi colori.

Per esempio, nei monitor televisivi a CRT si usa la sintesi additiva: i colori sono ottenuti miscelando la luce generata da sorgenti di luce rossa, verde e blu, costituite dai fosfori eccitati con l'intensità opportuna dal cannone elettronico. Nella stampa a colori si adotta invece la sintesi sottrattiva. Si usano degli inchiostri - nei sistemi tricromatici sono di colore giallo, ciano e magenta - che assorbono le componenti luminose diverse dal loro colore.

Dal punto di vista matematico la scelta dei primari è arbitraria. Nella pratica deriva dalla possibilità di ottenere le sorgenti luminose o i filtri assorbenti nei colori desiderati, considerando i parametri ottimali per la resa cromatica, per esempio la quantità di colori riproducibili (gamut).

Il principio della tricromia ha una giustificazione fisiologica. Infatti, i coni, recettori retinici attivi nella

Nota 15 - Si consideri che il segnale televisivo nasce sempre nel formato RGB poiché generato dai tre sensori inseriti nelle telecamere. La trasformazione nel formato YC_rC_b con relativo filtraggio è successiva, ma può essere realizzata già a livello di telecamera.
Nota 16 - Infitec™ è un acronimo per *interferenz filter technik* ed è un marchio di DaimlerChrysler Research and Technology di Ulm, di cui Barco (nota azienda di produzione di sistemi di visualizzazione) è licenziataria.

visione ad alti e medi livelli di luminosità [Nota 17] che consentono la percezione dei colori, sono differenziati in tre gruppi a seconda del tipo di sensibilità che presentano, dovuta alla presenza di particolari molecole chiamate fotopigmenti.

Gli andamenti della sensibilità dei coni in funzione della lunghezza d'onda della luce incidente, ossia del colore, sono riportati in figura 8.

Si noti che le curve di sensibilità presentano un valore di massimo che si trova a lunghezze d'onda differenti, in particolare nella zona del rosso (600 nm), del verde (550 nm) e del blu (450 nm). Per questo motivo è invalso l'uso di considerare i recettori come sensibili ai colori, rispettivamente, rosso, verde e blu. Ma questa è una semplificazione che non tiene conto che le curve di sensibilità dei recettori "rosso" e "verde" sono decisamente sovrapposte, e che, inoltre, i recettori "rossi" sono sensibili anche nella zona del blu.

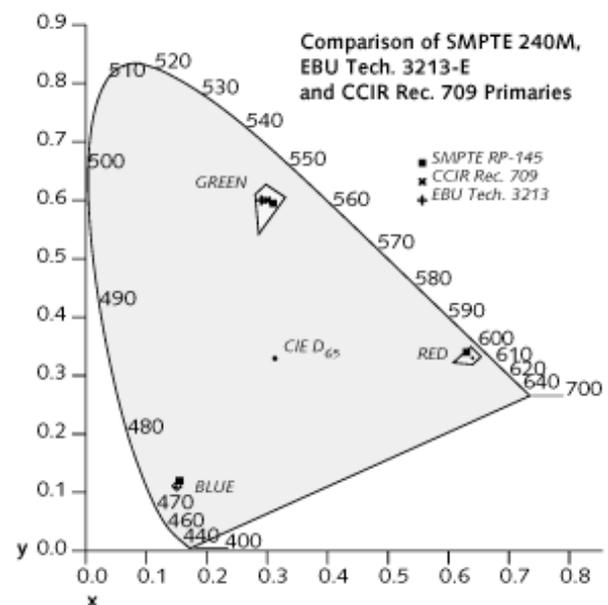
In ogni caso, gli apparati visualizzatori a colori, per esempio schermi a CRT o LCD, sono basati su terne di sorgenti luminose, una per ogni pixel, che emettono luce a lunghezze d'onda vicine a quelle di massima sensibilità retinica.

Fig. 8 – Andamenti della sensibilità dei coni in funzione della lunghezza d'onda della luce incidente.

Per ottenere omogeneità nella riproduzione dei colori, tali terne sono state standardizzate dai principali enti, per esempio ITU-R, EBU, SMPTE. In figura 9 si riporta la posizione dei primari stardardizzati dagli enti suindicati nel diagramma colorimetrico CIE; il primario rosso emette luce della lunghezza d'onda di circa 600-610 nm, il primario verde emette sui 550 nm e il primario blu emette luce a circa 470 nm, in accordo con gli andamenti di sensibilità dei fotorecettori.

Si noti che le diverse terne di primari non sono perfettamente coincidenti. Questo fa sì che alimentare un visualizzatore tarato su una terna colorimetrica con un segnale adatto per un'altra terna provoca una riproduzione colorimetrica non perfetta.

Nota 17 - Per completezza: la visione scotopica è la visione monocromatica dovuta unicamente all'attività dei bastoncelli della retina. Si tratta del tipo di visione che si ha quando il livello di illuminazione è molto basso e consente di rilevare differenze di brillanza ma non differenze di cromaticità. La visione a livelli di illuminazione normali è la visione fotopica, mentre quella a livelli intermedi è la visione mesopica.

Fig. 9 – Posizione dei primari colorimetrici ITU-R, EBU e SMPTE nel diagramma CIE. (fonte: [5])

Tuttavia la vicinanza tra i primari indica che il gamut sia pressoché lo stesso per le tre terne.

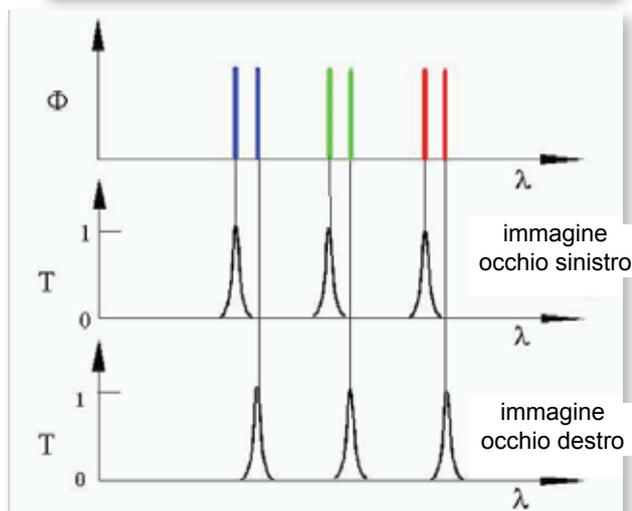
Analizzando gli andamenti di figura 8 si nota che l'intervallo di lunghezze d'onda attorno al massimo nel quale la sensibilità retinica si mantiene a valori elevati è relativamente ampio: circa 70 nm per il rosso, 80 nm per il verde e 45 nm per il blu.

Pertanto è possibile utilizzare terne di primari leggermente differenziati in lunghezza d'onda, purché nell'intorno del massimo della sensibilità dei recettori, per ottenere percezioni molto simili.

Questo fa sì che i diversi primari illustrati in figura 9 suscitino percezioni praticamente uguali; in altre parole, le tre terne colorimetriche standardizzate da ITU-R, EBU e SMPTE sono praticamente equivalenti.

3.2. Principio di funzionamento

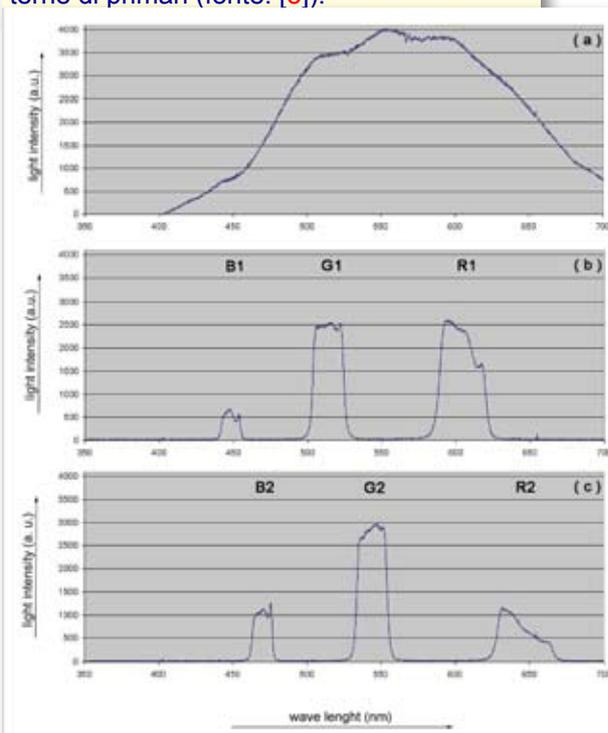
Il sistema Infitec sfrutta l'ampiezza delle curve di sensibilità retinica utilizzando due terne di primari colorimetrici molto vicini tra loro e posizionati nell'intorno del punto di massima sensibilità retinica, come schematizzato in figura 10 [5].

Fig. 10 – Coppia di terne di primari usati nel sistema Infitec (schematizzazione) (fonte: [5]).

Ciascuna terna colorimetrica viene associata ad un segnale della coppia stereoscopica che si vuole visualizzare - previo suo adattamento [Nota 18] per tenere conto della posizione dei nuovi primari colorimetrici rispetto a quelli utilizzati in fase di generazione.

Nei sistemi di visualizzazione a proiezione, le due terne di primari sono ottenute filtrando con filtri ottici di tipo passabanda la luce emessa da una sorgente luminosa che, in pratica, consiste in una lampada ad incandescenza del tipo utilizzato nei proiettori[Nota 19].

La figura 11 illustra lo spettro della sorgente e il risultato dei filtraggi passabanda.

Nota 18 - Si tratta di operare una trasformazione di base, in genere realizzata tramite prodotto matriciale con una matrice di trasformazione tra coordinate (corrispondenti alle terne di primari).
Nota 19 - Si ricorda che la Barco propone il sistema Infitec in sistemi stereoscopici a proiezione.

Fig. 11 – Spettro di una sorgente ad incandescenza (sopra) e il risultato dei filtraggi passabanda per ricavare le due terne di primari (fonte: [5]).

Per ottenere primari molto vicini in termini di lunghezza d'onda, le bande di transizione dei filtri devono essere piuttosto ripide. Per questo motivo si usano filtri dicroici che, a fronte di un maggior costo, presentano caratteristiche ottiche del livello necessario.

Nei sistemi a proiezione i due canali destro e sinistro vengono proiettati sovrapposti sullo stesso schermo, come schematizzato in figura 12; quest'ultimo non deve avere particolari caratteristiche ottiche a parte una buona capacità diffondente[Nota 20].

La separazione tra i due canali avviene a livello degli occhiali, illustrati in figura 13, indossati dall'osservatore ed equipaggiati con filtri con le stesse caratteristiche spettrali di quelli utilizzati per ricavare i primari colorimetrici.

Gli occhiali sono passivi. Tuttavia i filtri sono di materiale vetroso e quindi relativamente fragili e pesanti; inoltre sono a tutt'oggi relativamente cari.

I due canali della coppia stereoscopica vengono proiettati simultaneamente; non c'è sottocampionamento temporale, quindi il flicker non è apprezzabile, nel senso che rimane quello tipico del sistema video che alimenta i proiettori (nel caso televisivo europeo è a 50 Hz).

Il sistema Infitec ha prestazioni luminose paragonabili ai sistemi di proiezione standard 2D, in quanto la perdita di luminosità dovuta ai filtri è relativamente piccola, in particolare rispetto ai filtri polarizzatori, e non c'è sottocampionamento temporale come nei sistemi attivi.

3.3. Principali caratteristiche

Rispetto al sistema a proiezione di luce polarizzata, il sistema Infitec è caratterizzato da una migliore separazione tra canali dovuta alla elevata reiezione in banda attenuata delle componenti luminose indesiderate, caratteristica ottenuta anche grazie alle caratteristiche ottiche dei filtri dicroici[Nota 21] adottati. Pertanto il fenomeno del ghosting è talmente limitato da non risultare apprezzabile.

Nota 20 - Infatti, a differenza del sistema a proiezione di luce polarizzata, per il quale è necessario utilizzare uno schermo riflettente che mantenga la polarizzazione, nel sistema Infitec qualsiasi schermo per proiezione (che in genere sono di tipo diffondente) è adatto.

Nota 21 - I filtri dicroci o a riflessione sono fabbricati rivestendo uno strato di vetro con uno strato ottico. Questi filtri riflettono le porzioni di luce non volute alla sogente. Sono particolarmente indicati per lavori scientifici ad alta precisione, dato che la banda del filtro può essere selezionata con estrema precisione. Sono però molto più delicati e costosi dei filtri ad assorbimento.

Fig. 12 – Schema illustrativo di sistema di proiezione Infitec.

Due proiettori DLP o LCD

Un normale schermo

Un paio di occhiali stereo Infitec

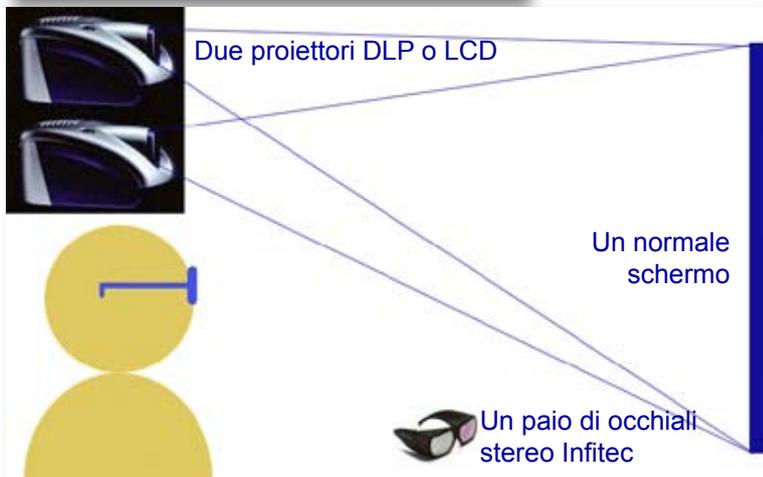



Fig. 13 – Occhiali con filtri Infitec (fonte: www.www.3dtiefenrausch.de).

La riproduzione colorimetrica è ottima.

Si noti che siccome la separazione dei canali viene effettuata nel campo delle frequenze, ossia dei colori, lo schermo non deve avere caratteristiche ottiche specifiche; in particolare non è necessario che sia di tipo riflettente, ma può essere un comune schermo per proiezione, con caratteristiche diffondenti. In questo modo si ottiene una migliore luminosità dell'immagine e un maggior angolo di visione rispetto alla tecnica di proiezione in luce polarizzata.

Bibliografia

1. G. Colace, M. Muratori – Televisione stereoscopica, le basi della tecnica stereoscopica – Elettronica e Telecomunicazioni, n. 2, Agosto 2004
2. E. Perucca – Dizionario di ingegneria - UTET
3. E. Dubois – A Projection Method to Generate Anaglyph Stereo Images – http://mti.xidian.edu.cn/multimedia/2001/supp/icassp2001/MAIN/papers/pap2222.pdf
4. Recommendation ITU BT.601-5 – STUDIO ENCODING PARAMETERS OF DIGITAL TELEVISION FOR STANDARD 4:3 AND WIDESCREEN 16:9 ASPECT RATIOS – ITU
5. H. Jorke, M. Fritz – Infitec - a new Stereoscopic Visualiztion Tool by Wavelength Multiplex Imaging – Proceedings Electronic Displays, September 2003, Wiesbaden (www.infitec.net/infitec_english.pdf)
6. C. A. Poynton – Wide Gamut Device-Independent Colout Image Interchange – Proceedings of the International Broadcasting Convention, 10-16 Sept. 1994, Conf. Publ. No. 397, copyright IEE, 1994 (www.poynton.com/PDFs/Colour_interchange.pdf)
7. M.Muratori – Tecniche per la visione stereoscopica – Elettronica e telecomunicazioni, n. 2, Aprile 2007.

OLOFONIA
una ripresa sonora di tutto ciò che ci circonda

dott. Leonardo **Scopece**
Rai
Centro Ricerche e
InnovazioneTecnologica
Torino

1. Introduzione

La storia dell'Uomo ha insegnato che si è sempre cercato di realizzare ciò che più è naturale per Lui, sfruttando l'ingegno di uomini che hanno dato spinte considerevoli nel mondo della scienza e della tecnologia.

Nel XIX Secolo, grazie a Hermann von Helmholtz e lord Rayleigh, si è studiata la propagazione delle onde e il comportamento dell'energia sonora; con l'invenzione del microfono e del fonografo nel 1877 è iniziata una nuova fase tecnologica con lo scopo di memorizzare su un supporto fisico eventi sonori che fanno parte del mondo che ci circonda. Col passare degli anni il microfono si è evoluto tecnologicamente fino ad arrivare a quel meraviglioso strumento di cattura, con caratteristiche meccaniche, elettriche e fisiche che tutti coloro che operano nel campo professionale possono apprezzare. Ma contemporaneamente si è cerca-

Sommario
Con l'avvento dell'alta definizione anche nel mondo del consumer e il conseguente formato 16/9 ormai consolidato, nasce l'esigenza di fornire un audio multicanale, all'altezza dell'immagine video.

I Broadcaster devono ormai attrezzarsi per cercare di produrre in modo sempre più economico ed efficace una traccia sonora multicanale, che meglio sposi l'esigenza di una visione delle immagini avvolgente ad una immagine sonora spaziale tridimensionale, che renda lo spettatore partecipe e all'interno della scena televisiva.

A questo scopo si ritiene riduttivo utilizzare tecniche di ripresa sonora stereofonica consolidate, le quali danno un risultato frontale e non tridimensionale.

Ecco perché il Centro Ricerche e Innovazione Tecnologica della Rai sta valutando e sperimentando nuovi sistemi di ripresa sonora che vadano incontro alle esigenze fin qui esposte, avendo come obiettivi l'ottimizzazione dei costi di produzione, una semplificazione dei problemi tecnici e ad un interfacciamento con le attuali infrastrutture di produzione audio. In questo articolo viene illustrato il sistema olofonico e le prime sperimentazioni effettuate.

to anche di utilizzarlo in modo tale da catturare nel modo più verosimile il mondo dei suoni da cui si è circondati.

Ecco quindi che nascono varie tecniche di ripresa sonora, dalla monofonia, quindi riproduzione di una sola sorgente sonora riprendendo con un solo microfono; alla stereofonia, termine che deriva dalla composizione di due parole greche stereo, solido, spaziale, e phōnía, voce, suono, che permette un ascolto sì più "aperto", più spaziale del sistema precedente, ma comunque limitato allo spazio compreso tra due altoparlanti frontali, riprendendo con almeno due microfoni; alla quadrifonia, un sistema stereofonico di registrazione e riproduzione contemporaneo su quattro canali. La ripresa quadrifonica si effettuava ponendo di fronte alla sorgente sonora quattro microfoni, registrando, ad esempio, su quattro tracce separate di un supporto magnetico e riproducendo tramite quattro altoparlanti che "avvolgevano" l'uditore. Ma il risultato ottenuto era abbastanza deludente rispetto alle aspettative, perché dava solo una riproduzione posteriore dei segnali frontali.

Da circa due decenni finalmente si riesce a riprodurre, con tecniche di simulazione fisica e algoritmi matematici, il suono nello spazio che ci circonda, utilizzando cinque o più altoparlanti.

2. Le tecniche per la ripresa stereofonica

Quando si vuole effettuare una ripresa sonora si hanno a disposizione varie tecniche di ripresa che, a parte quella monofonica, utilizzano più microfoni, o microfoni stereo con l'ausilio di microfoni "normali" per "rinforzare" particolari zone della scena, zone riprese fuori fuoco con i soli microfoni stereo.

Fino agli inizi degli anni '80 la tecnica più utilizzata dai broadcaster mondiali era senza dubbio la tecnica multimicrofonica. Essa consiste nella ripresa con "copertura" molto stretta, nel caso soprattutto di ripresa audio associata ad una ripresa video con la scenografia che ha la precedenza dal punto di vista artistico, e questo vale tuttora in programmi come il Festival di Sanremo, o nella ripresa di "zone orchestrali", che permettono di riprendere, ad esempio, i primi violini, i secondi violini, le viole, eccetera. E' chiaro che questa è l'unica tecnica che consente di avere il controllo pressoché totale sulla ripresa anche del singolo strumento in scena, basta aggiungere microfoni a sufficienza sul soggetto che si vuole riprendere. Ci sono vari problemi da risolvere prima della messa in onda o della registrazione: il posizionamento fisico dei microfoni, la loro messa in fase, il loro esatto posizionamento virtuale tramite il pan-pot, l'equilibrio energetico tra le varie sezioni o i vari strumenti ripresi. Il risultato è un suono molto "pulito", presente, completamente "a fuoco", un segnale stereo sicuramente mono-compatibile, ma con un fronte piatto, non profondo.

Allora si è cercato di ottenere un altro tipo di ripresa che desse la profondità di scena, ossia che permettesse di posizionare virtualmente gli strumenti musicali dove in realtà si trovano, sul palcoscenico. E' chiaro che a questo punto è nata l'esigenza di cambiare totalmente modo di "sentire": è preferibile avere tutti gli strumenti in primo piano, puliti e a fuoco, o dare all'ascoltatore la sensazione di realtà dell'evento sonoro? Cioè, quando si è davanti ad un'orchestra non si sentono in modo pulito tutti gli strumenti ma il cervello, grazie alla capacità di effettuare l'ascolto intenzionale è in grado, se vuole, di discriminare e esaltare alcuni elementi sonori o uno solo, in particolari condizioni di "pubblico educato". Quindi, quello che si è cercato di ottenere è un prodotto virtuale più vicino alla realtà, più "impastato", più "totale", senza possibilità, chiaramente di controllare totalmente le singole componenti delle sezioni sonore. Questo risultato si è ottenuto grazie a tecniche che hanno permesso di utilizzare non più semplici microfoni, ma singoli microfoni con doppia capsula a doppia membrana.

Le tecniche stereofoniche più utilizzate sono: l'XY, la Stereosonic, l'MS, la AB e la Testa artificiale. Queste tecniche appartengono a due

famiglie: tecniche a coincidenza, che utilizzano microfoni a doppia capsula presenti sullo stesso corpo microfono e in asse tra di loro, quindi a coincidenza di fase e con risultato sempre mono-compatibile; tecniche ad esclusione, con microfoni che hanno le due capsule distanziate nello spazio, l'AB e la Testa artificiale, ossia non in asse tra loro e con la possibilità che l'onda sonora, se arriva lateralmente al microfono, compia percorsi diversi per arrivare alle due capsule, quindi con fasi differenti, e di conseguenza non sempre il segnale ottenuto risulta mono-compatibile.

La tecnica XY utilizza un microfono a coincidenza. Le capsule sui microfoni a coincidenza sono tali per cui una rimane rigidamente ferma rispetto al corpo microfono, l'altra la si può ruotare con un angolo anche abbastanza ampio in tutta la sua rotazione e con questa tecnica, come per le altre che si stanno analizzando, l'angolo tra le capsule è quello canonico di 90°.

Ad esempio, il microfono Neumann SM 69 fet ha la capsula non fissa che può ruotare di 270° e per ognuna delle due capsule si può scegliere la direttività in modo indipendente tra omni-direzionale, cardioide o bi-direzionale.

Il risultato acustico della tecnica XY non è pienamente soddisfacente perché in ascolto il fronte dà l'impressione di essere abbastanza stretto, gli elementi S1 ed S7 (figura 1) non risultano uscire dai due altoparlanti posti in modo canonico a ±30° rispetto all'osservatore, ma formano un angolo di circa ±10°.

La tecnica Stereosonic, sempre con un microfono a coincidenza, ha ambedue le capsule selezionate con direttività a 8 o bi-direzionale e ruotate una rispetto all'altra di 90°. Il fronte sonoro in questo caso è ampio, va dall'altoparlante di sinistra a quello di destra, ma si verificano un paio di problemi: la sorgente S1 (figura 2) viene ripresa con una sensibilità maggiore rispetto ad S2 ed S3, risultando così rientrante al centro rispetto al fronte, quindi questo è arcuato e rientrato al centro; le capsule posteriori è vero che sono in controfase rispetto a quelle anteriori ma sono in fase tra loro, quindi c'è uguale sensibilità tra la coppia di lobi posteriore e la coppia di lobi anteriore, e non si deve quindi posizionare il microfono in verticale, ma rivolto verso la scena sonora sperando che non vi sia molto rumore di ambiente.

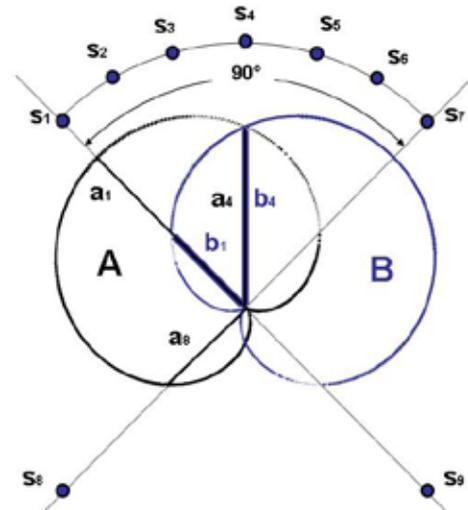



Fig. 1 – Tecnica XY.

Fig. 2 – Tecnica Stereosonic.

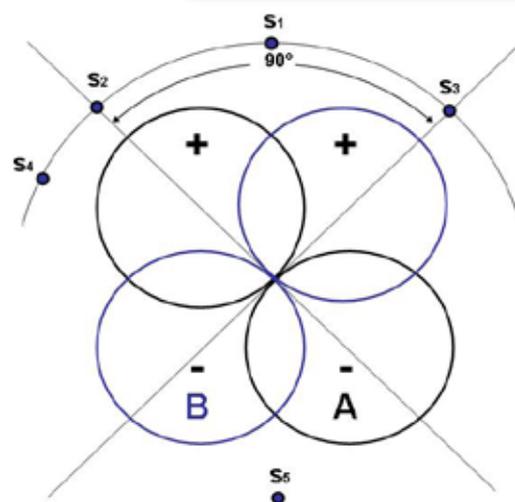
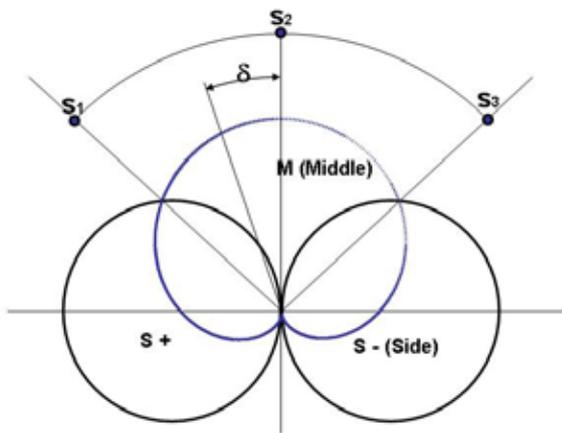



Fig. 3 – Tecnica MS.

La tecnica MS, l'ultima a coincidenza che si considera, consiste nel selezionare una capsula in situazione bi-direzionale e l'altra in modo cardioide, questa puntata verso il palcoscenico, e a 90° una dall'altra (figura 3). Il segnale dell'8 viene sdoppiato e inviato su due ingressi di un mixer, messi in controfase uno rispetto all'altro, ottenendo così due segnali che chiamiamo –S e +S. Il segnale dell'altra capsula, che chiamiamo M, viene inviato ad un terzo ingresso del banco mixer.

Quando si miscelano i tre dosatori si ottengono le somme (M+S) ed (M-S), che corrispondono di fatto alla decodifica stereo che dà luogo ai segnali Left e Right. Quindi questa tecnica, più che di ripresa stereo, la si può considerare di codifica stereo. Il risultato acustico è molto buono, ampio, profondo ma con una particolarità: le alte frequenze ad un livello un po' sostenuto sembra che "svolazzino", sembra che gli strumenti che le generano non siano localizzabili in modo preciso, ma è sicuramente un prezzo accettabilissimo da pagare in cambio della qualità che si ottiene.

Fig. 4 – Microfono Neumann KU 100.

La tecnica a Testa artificiale è stata introdotta in Germania. Si pensò che per poter effettuare una ripresa, la più reale possibile, bisognasse ricreare in campo di ripresa le stesse condizioni fisiche della presenza umana, quindi riflessione e diffrazione della testa dell'uomo. All'interno dei lobi auricolari (figura 4) sono inseriti due microfoni omni-direzionali che rendono questa tecnica ad esclusione. Il risultato è buono per l'ascolto in cuffia, molto meno per ascolto in aria, ossia su altoparlanti, probabilmente perché per l'ascolto in aria, oltre la riflessione e la diffrazione della Testa artificiale in fase di ripresa, gli stessi fenomeni sono generati anche dalla testa dell'osservatore, in fase di ascolto.

In cuffia si ha un ascolto più piacevole di quello che si può ottenere con altre tecniche, perché l'immagine sonora, invece di concentrarsi lungo la striscia che unisce i due padiglioni del mezzo trasduttore, e quindi dentro la massa cranica, si "visualizza" all'esterno della testa, rendendo l'ascolto più piacevole e meno stressante.

Infine consideriamo la Tecnica AB. Anche questa è ad esclusione ed è stata studiata e realizzata in Francia.

Qui si è sfruttato il discorso accennato prima e si è evitata la doppia riflessione e la doppia diffrazione, non ponendo più una massa fisica tra i microfoni. Si è invece costruito un supporto per due capsule cardiodi poste a 17 cm di distanza, come mediamente sono poste le membrane timpaniche nell'orecchio umano e con un'angolazione di 110° una dall'altra, inclinazione media dei padiglioni auricolari (figure 5 e 6).

Il risultato è decisamente buono, spaziale, dà una profondità della scena molto reale e, timbricamente, il suono è ricco e gradevole.

E' chiaro che tutte queste tecniche di ripresa stereo, per un palcoscenico ricco di strumenti, non possono avere tutti i suoni a fuoco, e per ovviare al problema ci si affida a microfoni detti di supporto o di rinforzo che vengono posizionati sulle sorgenti sonore sfocate, microfoni che bisogna dosare in modo opportuno rispetto al risultato della tecnica stereo di base e alcune volte bisogna anche inserire un ritardo sul percorso del loro segnale.

Fino qui si sono considerate le tecniche di ripresa stereofoniche convenzionali, ma non sono riprese da cui si ottengono suoni veramente spaziali, totali, perché riproducono virtualmente le sorgenti sonore in modo frontale, con un fronte più o meno aperto e più o meno profondo.

3. La tecnica Olofonica - Teoria

Agli inizi degli anni '80 il ricercatore italo-argentino Hugo Zucarelli ha applicato il modello olografico ai fenomeni acustici, incuriosito dal fatto che gli uomini sono in grado di localizzare la sorgente sonora senza indirizzare l'attenzione verso essa, e continuano ad avere questa capacità anche se sordi da un orecchio. Le sue conclusioni furono poi sviluppate e portarono ad un brevetto da parte di Umberto Gabriele Maggi, che insieme a suo figlio Maurizio, realizzò nel 1983 l'"holophone", uno speciale microfono capace di riprodurre suoni in tre dimensioni. Il primo album realizzato con questa tecnica fu "Final Cut" dei Pink Floyd, per il quale Maurizio Maggi fece il tecnico audio. Al tempo non diede buoni risultati se non in cuffia, dove i suoni avvolgevano letteralmente l'osservatore e la sensazione della realtà era impressionante. In aria l'effetto non era affatto accettabile e solo negli ultimi anni la tecnologia ha portato alla realizzazione di microfoni che hanno permesso di ottenere un risultato molto buono anche con ascolto in aria.

Fig. 5 – Tecnica AB.

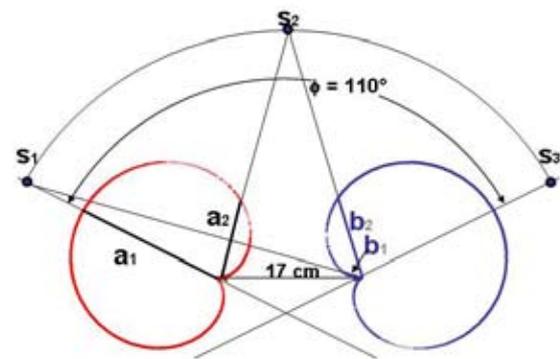
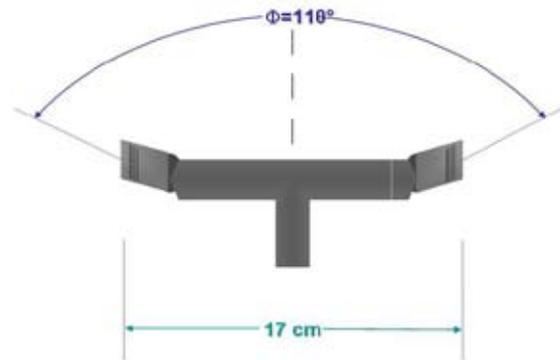



Fig. 6 – Figura polare della tecnica AB.

L'etimologia della parola olofonia deriva dalle parole greche hólos, tutto, e phōnía, voce, suono. Ma su quali principi si basa la tecnica olofonica?

I microfoni moderni Holophone si basano sulla teoria dell'HRTF (Head Related Transfer Function), che è a completamento delle ricerche effettuate sulla teoria Duplex.

La teoria Duplex si fonda sulla stima del fenomeno della localizzazione spaziale di eventi sonori da parte delle orecchie umane. Gli effetti principali alla base di questo fenomeno sono: l'ITD (Interaural Time Difference) e l'IID (Interaural Intensity Differences).

Per comprendere in cosa consistono i due effetti appena enunciati si osservino le figure 7 e 8.

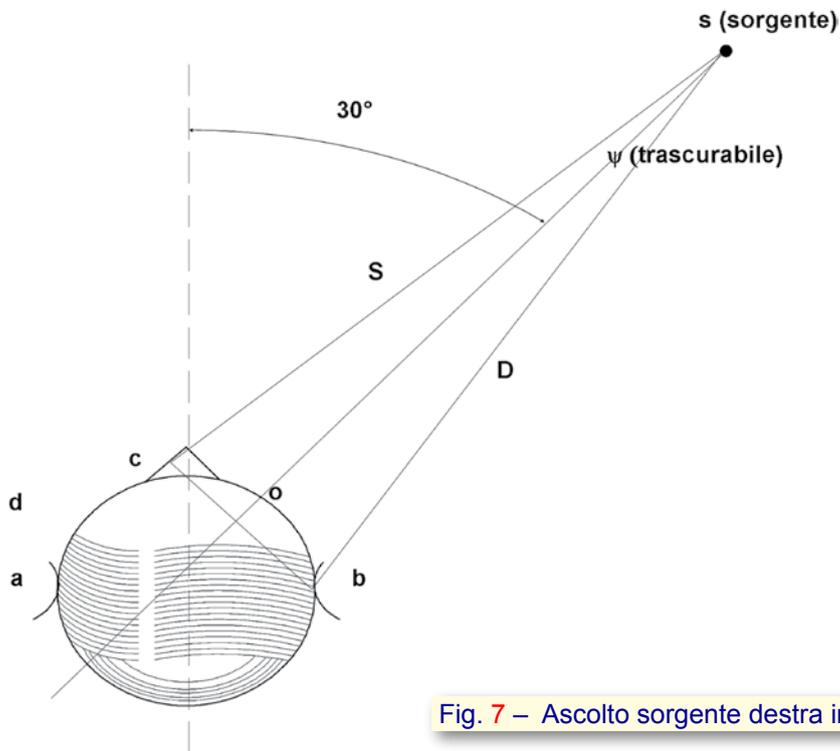



Fig. 7 – Ascolto sorgente destra in condizione stereo canonica.

Si inizi con l'ITD. Si consideri una delle due sorgenti sonore nell'ascolto stereofonico, la destra, e la si consideri abbastanza distante dall'osservatore, in *campo lontano*, tale per cui l'onda generata, che sia piana o sferica, non abbia più importanza e la si possa considerare solo piana. Dalla figura 7, quindi, si deduce che l'angolo sotteso al vertice del triangolo isoscele con base *cb,* con cui si rappresenta l'onda piana che arriva all'osservatore, sia trascurabile. Inoltre si ipotizzi la presenza della testa dell'osservatore con una sfera di diametro di 17 cm.

Quando l'onda sonora colpisce l'osservatore ha "colpito" in *b* l'orecchio destro, ma l'altro vertice *c* del triangolo non è ancora arrivato all'orecchio sinistro. Tenendo conto che la velocità dell'onda sonora in aria si può approssimare a circa 340 m/s, si è calcolato che per percorrere la distanza *ca* di 8,5 cm l'onda impiega un tempo di 250 μs, che potrebbe sembrare ininfluente in ascolto, ma come differenza binaurale l'uomo riesce a percepire ritardi di tempo anche inferiori, di 10÷20 μs, di pochi gradi di sfasamento, che significa percepire il cambio apparente della direzione della sorgente sonora.

Con semplici prove si può dimostrare che per frequenze inferiori a 1500 Hz, per cui la lunghezza d'onda diventa paragonabile con la distanza tra le orecchie, l'orecchio destro, nell'esempio visto in precedenza, percepisce l'onda sonora in anticipo rispetto all'orecchio sinistro, e l'uditore percepisce chiaramente la posizione spaziale della sorgente, con un preciso angolo di azimuth. Nel caso si analizzino invece frequenze uguali o maggiori a 1500 Hz, il ritardo interaurale porta prima ad una ambiguità e poi a problemi veri e propri di localizzazione spaziale della sorgente sonora.

Quindi, per queste frequenze diventa importante la percezione della differenza energetica dell'onda tra un orecchio e l'altro, e subentra come effetto l'IID. La differenza di intensità interaurale diventa fondamentale per la esatta localizzazione spaziale di frequenze al di sopra dei 1500 Hz.

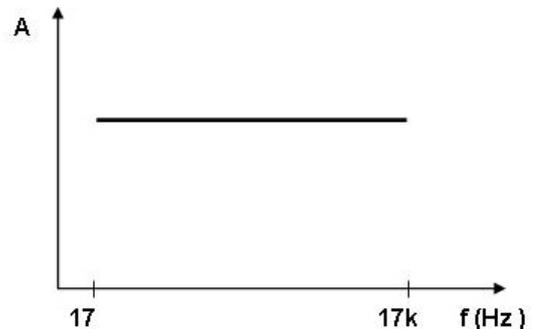
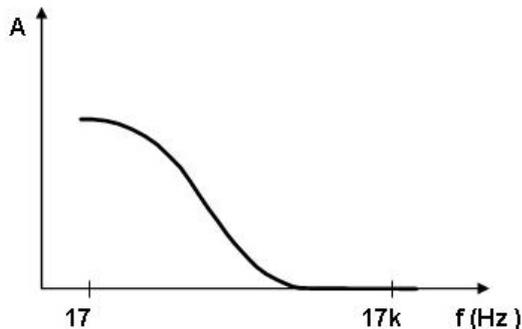



Fig. 8 – Differenza di ampiezza del segnale da destra in ascolto stereo canonico.

Si consideri una sorgente sonora in campo libero, senza pareti. In queste condizioni ciò che può influenzare il sistema di analisi è il corpo umano: la testa, con le orecchie e il naso. Infatti, quando una sorgente è lontana, colpisce la testa e si deve tenere conto delle dimensioni della testa stessa, e da che direzione arriva l'onda sonora. Ricordando il campo di frequenze udibili, da circa 20 Hz a circa 20000 Hz, e conseguenti lunghezze d'onda in gioco, da 17 m a 17 cm, si capisce come un corpo in presenza dell'onda possa influenzare il percorso dell'onda stessa, a causa dei fenomeni della diffrazione, della riflessione e del decadimento di energia in ragione di $1/R^2$. Si supponga che la sorgente sonora si trovi sullo stesso piano della testa ma a 90° dalla posizione frontale, massima ricezione da parte di un orecchio e minima dall'altro; si supponga che la differenza di percorso dell'onda, da quando ha colpito un orecchio all'altro, sia di circa 25 cm. La differenza di livello di pressione sonora tra un orecchio e l'altro è di solo 0,4 dB, che si può considerare impercettibile. Se la sorgente è molto vicina alla testa ma in posizione casuale e non frontale, diventa sostanziale la differenza di livello tra un orecchio e l'altro. Inoltre la testa lavora come corpo mascherante per una certa gamma di frequenze. Ciò porta al fatto che un segnale con un certo equilibrio timbrico su un orecchio, quello più vicino alla sorgente sonora, non è lo stesso che percepirà l'altro orecchio (figura 8).

La figura 9 mostra la differenza di ampiezza misurata da Steinberg e Snow nel 1934, risultato che si basa su test effettuati con frequenze pure nel 1931 da Sivian e White.

Fig. 9 – Differenza di ampiezza in funzione delle frequenze.

La teoria Duplex spiega però solo la capacità di localizzazione spaziale della sorgente se questa si trova sullo stesso piano dell'osservatore, sul piano azimutale. Ma con questa teoria, se si tenta di localizzare una sorgente sonora sul piano mediano, ossia con angolo di elevazione differente rispetto alla testa, è possibile arrivare a percepire infiniti punti alla stessa distanza dall'orecchio con stesso ITD e IID, creando ciò che viene definito *cono di confusione* (figura 10).

Dal momento che si può capire la provenienza di un'onda nello spazio e non solo nel piano azimutale, si è ipotizzato che questa capacità sia dovuta ad un meccanismo di ascolto monoaurale, che tiene conto della risposta spettrale dell'evento sonoro da parte del canale uditivo e dalla conformazione del padiglione auricolare.

Questo ragionamento è complicato da formalizzare matematicamente. Le ricerche che si sono susseguite nel corso degli anni per comprendere come si possa localizzare una sorgente sonora nello spazio, hanno portato ad un data-base delle cosiddette Head Related Transfer Function (HRTF), funzioni che tengono conto di come il suono proveniente da vari punti dello spazio viene filtrato dall'orecchio umano. Cioè, si sono registrate le risposte impulsive dell'orecchio ad eventi generati in vari punti dello spazio con varie angolazioni, per poi analizzarle e cercare di capire quali possano essere le relazioni che regolano questi risultati.

Le HRTF sono funzioni che, con opportuni ritardi introdotti in un segnale sonoro, permettono di capire la sua localizzazione spaziale. Per poter calcolare, quindi, la pressione acustica dovuta ad una sorgente qualsiasi x(t) nello spazio, bisogna conoscere la risposta all'impulso del timpano, la cui funzione è definita Head Related Impulse Response (HRIR). La trasformata di Fourier H(f) di tale risposta è appunto la HRTF.

Per elaborare un suono tridimensionale tenendo conto, quindi, dei fenomeni prima elencati, cioè l'ITD, l'IID e la modellazione della struttura spettrale a causa delle riflessioni, diffrazioni e conformazione del padiglione auricolare, si utilizza una coppia di filtri FIR (Finite Inpulse Response) di lunghezza appropriata. Con i FIR un segnale sinusoidale che li attraversa uscirà ancora sinusoidale, ma scalato in ampiezza e ritardato in fase in funzione della sua risposta in frequenza. Per ridurre la quantità di calcolo di detto filtro, che può arrivare ad essere dell'ordine di 128 punti, alcuni autori consigliano il filtro IIR (Infinite Impulse Response) a basso ordine, generalmente di 6÷12 punti. A parità di ordine, a differenza dei filtri FIR, i filtri IIR permettono risposte in frequenza più ripide, introducendo però sempre delle distorsioni di fase, ossia le componenti spettrali vengono ritardate in funzione della frequenza.

Le HRTF sono misurate su entrambe le "orecchie" poste su un manichino, fissando un angolo di riferimento rispetto alla testa, e tali funzioni sono misurate con sorgenti poste a diversi angoli azimutali e diversi angoli mediani, di elevazione. Le HRTF includono anche le informazioni di differenza di tempo interaurale e di differenza di intensità interaurale. Le ITD sono comprese nello spettro di fase del filtro FIR, mentre le IID sono comprese nella risposta in potenza del FIR.

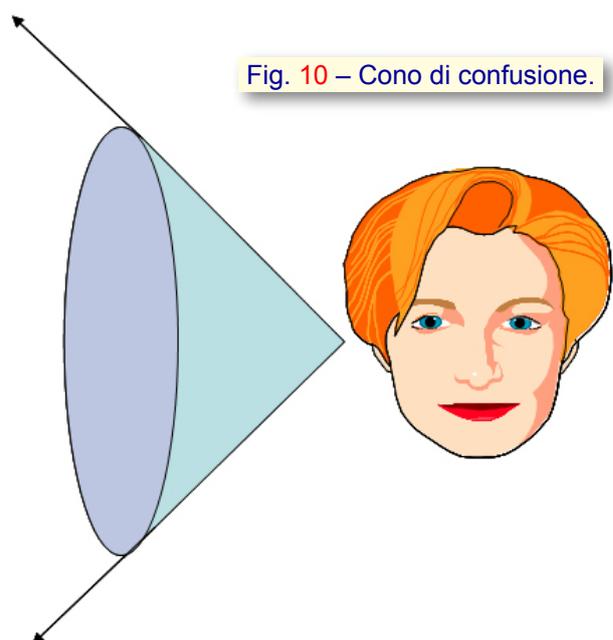



Fig. 10 – Cono di confusione.

Tornando al manichino, si utilizzano due microfoni inseriti nei suoi padiglioni auricolari e si inviano, tramite un altoparlante posto con un certo angolo di azimuth, un certo angolo di elevazione e ad una certa distanza, diversi stimoli che potrebbero essere anche dei semplici click o sequenze binarie pseudo-random. La funzione di trasferimento che si ottiene comprende, ovviamente, anche le funzioni di trasferimento degli apparati utilizzati per la misura, come i microfoni e gli altoparlanti. Queste funzioni di trasferimento sono chiamate Common Transfer Function (CTF) e vengono poi eliminate dalla misura della risposta complessiva del sistema.

Come risultato finale delle misure si ottiene un set di funzioni di trasferimento direzionale, le Directional Transfer Function (DTF), ognuna per una ben precisa posizione nello spazio. E' proprio questo set di funzioni che viene chiamato HRTF.

Le tecniche di rilevazione e di calcolo delle HRTF si basano su vari metodi, ma i due principali sono i seguenti.

Il primo con modelli sferici che simulano la testa umana. In questo caso, utilizzando una sorgente sonora piana ad una data frequenza, si può calcolare la pressione acustica su due punti della sfera dove si ipotizza siano localizzate le orecchie. Si effettuano vari calcoli a differenti frequenze e a diversi angoli azimutali e mediani, ottenendo così un set di HRTF.

Per il secondo modello si utilizza un manichino con due microfoni nei lobi auricolari. Si effettuano le misure in camera anecoica e con un gran numero di punti di generazione sonora nello spazio.

4. La tecnica Olofonica – Sperimentazione

Vari costruttori, da un po' di anni, hanno cominciato a soddisfare l'esigenza da parte dei broadcaster di poter effettuare riprese sonore che possano dare un risultato sonoro immediato dell'ambiente. Tra i prodotti che hanno questo scopo, il Centro Ricerche della Rai ha scelto di valutare e sperimentare un microfono della ditta canadese Holophone, che è stato utilizzato in trasmissione dalla Fox, NBC, ABC, CBS, CBC, e ESPN, oltre che dalla EA Sports per videogiochi, e da artisti famosi quali Elton John, Celine Dion e Iron Maiden e per riprese sportive (NBA Basketball, NFL Football e NHL Hockey).

La ditta Holophone produce tre modelli di microfoni olofonici: H2-PRO, H3D e H4 SuperMini.

L'Holophone H2-PRO (figura 11), scelto per questa sperimentazione, è un microfono professionale che ha sette capsule localizzate esternamente sul perimetro del supporto definito *testa*, e una interna che cattura la bassa frequenza. Può ottenere come risultato 5.1, 6.1 o 7.1 canali con suono surround. Tutti i suoni registrati tramite questo microfono sono discreti e succedono in tempo reale senza necessità di manipolazione da parte di mixer o quant'altro e ciò lo rende molto flessibile per l'utilizzo in studio, per la registrazione in formati surround e per tutti i formati consumer come Dolby, DTS e Circle Surround.

Fig. 11 – Holophone H2 PRO.

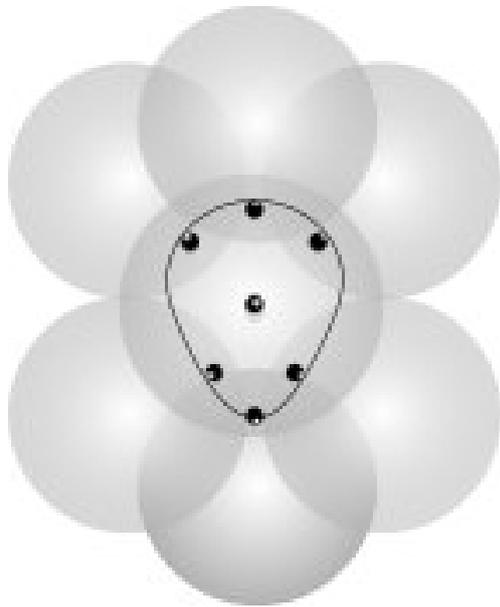



Fig. 12 – Diagramma polare dell'Holophone H2 PRO.

Le sette capsule esterne, come si è detto, sono poste sulla superficie esterna della testa: sono capsule omni-direzionali, modello "4060" della ditta danese DPA, che danno la configurazione direzionale, montati sull'H2 Pro, della figura 12.

Le sette capsule montate esternamente hanno un range di frequenza da 20 Hz a 20 kHz, mentre quella interna da 20 Hz a 110 Hz. La sensibilità è di 20 mV/Pa a 1 kHz con livello di pressione massima pari a 134 dB_{SPL} prima del clipping.

Il suono surround, così, dovrebbe essere riprodotto abbastanza fedelmente, anche prendendo le uscite dirette dal microfono e inviandole a degli altoparlanti.

Si ricordi brevemente il percorso storico per arrivare ad ottenere un suono veramente spaziale e totale. Nel 1940 Disney introdusse il suono surround nei cinema in occasione della sua produzione "Fantasia", utilizzando tre speaker dietro lo schermo e altri posti in posizione posteriore. Nel 1950 prese piede la registrazione stereofonica che parte dal presupposto che si ascolta con due orecchie. Poi si ebbero delle prove sulla quadrifonia, che non prese piede per assenza di materiale, costi elevati dei sistemi e pochissimo mercato. Nel 1970, con "Star Wars", George Lucas introdusse il Dolby Stereo che fu poi matricizzato portando ai quattro canali left, right, center e rear. Oggi si hanno sistemi Dolby Digital che impiegano sei sorgenti: center, left, right, left surround, right surround, LFE (Low Frequency Effects). Detta configurazione è conosciuta come 5.1. Un sistema in competizione con il Dolby è il DTS (Digital Theater Systems) che fu introdotto in occasione del film "Jurassic Park". Altro sistema che supporta il 5.1, è il Circle Surround. L'IMAX usa il 6.1 aggiungendo un canale superiore al 5.1, configurazione supportata anche da Holophone. Solo recentemente Dolby, DTS e SRS Labs hanno esteso i loro sistemi alla configurazione 6.1 aggiungendo un canale posteriore centrale (Dolby Digital EX, DTS ES e Circle Surround II). I nfine la Sony ha introdotto un nuovo standard, SDDS (Sony Dynamic Digital Sound) con la configurazione 7.1, che ha tolto il canale superiore e ha introdotto il left center e il right center.

L'Holophone H2 PRO in uscita ha otto cavi bilanciati XLR così assegnati:

Canali	Microfono Holophone H2 PRO
1	Left
2	Right
3	Center
4	LFE
5	Left Surround
6	Right Surround
7	Top
8	Center Rear

Si illustrano ora le sperimentazioni effettuate dal Centro Ricerche con il microfono olofonico.

La prima sperimentazione aveva lo scopo di verificare la "pulizia" dei suoni degli strumenti durante l'esecuzione di un'orchestra sinfonica.

Presso l'Auditorium Rai di Torino, e con la collaborazione dei tecnici di Radiofonia, si è posto il microfono su un'asta all'altezza di circa 3,5 metri, rivolto verso il centro del palcoscenico, alle spalle del direttore d'orchestra (figura 13).

Le otto uscite del microfono sono state collegate al mixer (figura 14) che alimentava con la phantom le capsule a condensatore del microfono e preamplificava i segnali prima di inviarli in registrazione, singolarmente, in un registratore digitale (figura 15).

Una volta ottenuta la registrazione su otto tracce separate del nastro digitale, si sono acquisite tali tracce tramite un Macintosh PowerPC G5, con quattro processori da 2,5 GHz cadauno, DDR2 SDRAM da 4,5 GB e velocità di Bus da 1,25 GHz. Con il programma "Soundtrack" si sono poi indirizzati i segnali sui 5.1 speaker in regia.

Fig. 13 – Ripresa all'Auditorium Rai di Torino durante le prove della Sinfonia n. 2 in mi minore op. 27 di Sergej Rachmaninov, diretta da Kwamé Ryan.

Fig. 14 – Mixer Yamaha 01V 96.

Fig. 15 – Registratore digitale Tascam DA-98 HR.

Il risultato ottenuto è qualcosa di inaspettato. Ci si attende un suono frontale, con pochissimo effetto surround. Riascoltando il risultato, invece, sembra di essere immersi all'interno dell'orchestra, circondati da essa, e i timbri dei vari strumenti sono abbastanza reali.

La seconda sperimentazione effettuata, è consistita nella ripresa di effetti allo Stadio Olimpico di Torino. La ripresa sonora è stata affiancata dalla ripresa video in alta definizione (figura 16).

Fig. 16 – Riprese allo Stadio Olimpico di Torino durante la partita di campionato tra Juventus e Mantova. Si notino la *testa* olofonica e la telecamera Sony XDCAM HD PDW-F350L.

In questa occasione, la ripresa è stata effettuata alle spalle di una delle due porte, leggermente spostata a destra rispetto alla porta stessa, sotto la curva che ospitava i tifosi juventini. Le condizioni di ripresa erano vincolate dal permesso concesso dagli organizzatori.

Per poter registrare l'audio in telecamera, si è usato un codificatore (figura 17) che permette di codificare sei canali del microfono: si è fatta la scelta di escludere il center rear e il top, e le uscite si sono inviate come Total Left e Total Right, codificate, ai due canali esterni della telecamera come segnali stereo.

Una volta in regia, i sei canali audio sono stati "riestratti" decodificandoli con il decoder (figura 18). Sia il codificatore che il decodificatore operano con il sistema Circle Surround.

Il risultato, in questo caso, è stato stupefacente per l'impressione che si è avuta sulla realtà degli effetti spaziali riprodotti.

Fig. 17 – Codificatore CSE-06P della SRS Labs.

Fig. 18 – Decodificatore CSD-07 della SRS Labs.

5. Future Sperimentazioni

Il risultato ottenuto può essere definito inaspettato e originale, per quanto riguarda la ripresa musicale, così come si rivelò il passaggio dalla ripresa con tecnica multimicrofonica, con tutti gli strumenti presenti e a fuoco, alle tecniche stereo. L'obiettivo allora era di cambiare filosofia di ascolto, ossia dare un risultato più veritiero dell'evento sonoro: non si puntava ad una maggiore pulizia degli strumenti, bensì ad un maggior realismo dei suoni, come in presenza dell'orchestra.

Oggi, nella sperimentazione effettuata, si rivoluziona in qualche modo il tipo di ascolto, si è immersi nel cuore dell'orchestra, circondati dagli strumenti.

Per quanto riguarda la ripresa allo stadio, invece, si è ottenuto il risultato previsto, di realtà sonora spaziale.

Si intende proseguire la sperimentazione con una serie di riprese, sempre di musica classica, e pertanto sono stati presi contatti preliminari con il Centro di Produzione Rai di Milano.

Individuando posizioni opportune per il microfono, più in alto sulla scena e/o più lontano dal direttore d'orchestra, si verificherà se il risultato è conforme alle aspettative, cioè un risultato frontale, con un po' di effetto posteriore dovuto al riverbero dell'ambiente. Per quanto riguarda la "pulizia" degli strumenti, l'obiettivo è quello di far valutare i risultati a musicisti e ingegneri del suono, esperti atti a giudicare questo parametro molto soggettivo. Si ritiene che, anche con questa tecnica come per quelle stereo, se l'orchestra è disposta in uno spazio abbastanza ampio, siano necessari microfoni di "supporto" sulle sorgenti sonore distanti e leggermente sfuocate, da mixare in modo opportuno con le uscite dell'Holophone.

Sarà interessante effettuare un'altra ripresa allo stadio di calcio ponendosi questa volta, se possibile, in posizione centrale rispetto al campo e provando a "inseguire" con il microfono l'azione, seguendo i movimenti della telecamera. Questa tecnica è interessante perché non complica la vita agli addetti ai lavori e perché può essere utilizzata la stessa infrastruttura già esistente negli stadi, cavi e eventuale mixer: l'Holofone H2 Pro può essere posizionato facilmente in campo, su un semplice stativo. Comunque, se si vogliono riprendere gli effetti del colpo sul pallone o del palo colpito, è chiaro che oltre questo microfono converrà lasciarne in campo altri, ed è certamente indubbio il grande risparmio logistico e operativo per la ripresa totale degli effetti.

Altro evento sportivo di interesse, è una corsa ciclistica, seguendo gli atleti su un mezzo motorizzato, per sentire le voci e le grida del pubblico che "corrono" ai lati della strada mentre si avanza.

Altre possibili applicazioni sono per News e riprese in uno studio televisivo, anche se si prevede sia necessario comunque microfonare opportunamente i partecipanti.

In base ai primi risultati della sperimentazione, si ritiene che in molti casi con questa tecnica si possa ottenere un risultato sonoro spaziale, da utilizzare direttamente sugli impianti 5.1 o superiori.

L'avvento del digitale ha consentito al broadcaster di fornire all'utente servizi migliori e più coinvolgenti, sia in campo audio che video. Questo comporta cambiamenti significativi anche nella catena di produzione e nella relativa organizzazione del lavoro, con conseguenze sui costi.

La tecnica olofonica potrebbe contribuire ad ottenere risultati ottimali dal punto di vista della ripresa e dell'ascolto, con un minimo impatto sugli investimenti e la riorganizzazione delle risorse e dell'utilizzo delle infrastrutture tecniche già esistenti.

Bibliografia

1. L. Scopece: “L’audio per la televisione. Manuale di audio - 1° Vol.” – Gremese Editore, Roma 2008
2. B. Agostino: “Sistemi di ripresa per il surround sound” – Tesi al Politecnico di Torino, luglio 2006
3. R. Prosperi: “Elementi di acustica e stereofonia 2° Vol.” - Ed. KLIM, Roma 1987
4. L. Scopece: “Cenni di Audio” - Pubblicazione interna Rai, Roma 1996
5. E. F. Alton: “Manuale di acustica” - Ed. Hoepli, Milano 2000
6. Glen M. Ballou: “Handbook for Sound Engineers” - Editor, USA 1998
7. “Audio Engineer’s Reference Book” - Edited by Michael Talbot-Smith, Great Britain 1994
8. L. Scopece, B. Agostino:”Nuove tecniche di ripresa Sonora in surround” – Relazione tecnica di Rai - Centro Ricerche, 2006
9. http://www.holophone.com
10. http://www.srslabs.com/pressarticle322.asp?sid=167
11. http://www.marcostefanelli.com/olofonia/suonoemo.htm
12. http://xoomer.alice.it/e.giordani/docs/spat_a.PDF
13. http://giorgio.cmlug.org/Audio3d/Tesina/cap1.html
14. http://www.vialattea.net/esperti/php/risposta.php?num=8783
15. http://cnx.org/content/m12827/latest/

Che cosa è, come funziona:

Alta definizione: display 1080p

ing. Marzio **Barbero** e
ing. Natasha **Shpuza**

Sommario
Precedenti articoli sono stati dedicati all'evoluzione dei display utilizzati per la TV e la HDTV. Fra le tecnologie ormai mature vi sono gli schermi al plasma, LCD-TFT, a retroproiezione. La competizione, la conseguente riduzione dei prezzi e dei margini di profitto, spingono i produttori a realizzare schermi con caratteristiche innovative e attraenti per i consumatori. La piena definizione, 1920 pixel x 1080 righe, è ormai disponibile in prodotti basati su tutte le tecnologie precedentemente citate e l'associazione dell'industria europea EICTA ha recentemente definito i loghi "HD ready 1080p" e "HD TV 1080p". La tendenza a sviluppare display a più basso consumo, di minor peso e soprattutto più sottili, è accelerata dagli annunci sulla imminente produzione di schemi TV basati su composti organici elettroluminescenti (OLED).

1. Cresce la definizione e diminuisce lo spessore

Un articolo di due anni fa [1] offriva una panoramica sulle tecnologie alla base degli schermi piatti ed evidenziava il successo di questi prodotti, dovuto ad impianti di produzione sempre più efficienti con un conseguente miglioramento delle caratteristiche e una rapida diminuzione dei costi. Allora, come oggi, forte era la competizione delle industrie giapponesi e coreane, che annunciano, e puntualmente mettono in produzione, schermi al plasma (PDP) e a cristalli liquidi (LCD) con caratteristiche innovative.

Nei due anni trascorsi il mercato degli schermi piatti per TV è stato in crescita in termini di volumi (Tabella 1), ma alla crescita in termini qualitativi e quantitativi ormai non corrisponde più un andamento altrettanto favorevole dei margini di profitto per le industrie del settore.

Tab. 1 - Forniture delle industrie giapponesi nel 2006 e previsioni per il 2007 in migliaia di pezzi (fonte: www.displaybank.com)

	2006		2007	
	LCD TV	PDP TV	LCD TV	PDP TV
Sharp	6030		9000	
Sony	6300		10000	
Matsushita	2400	3500	4000	5000
Hitachi	510	770	800	1400

Le industrie quindi si impegnano nella ricerca di nuovi mercati, in aggiunta a quelli per PC e notebook, schermi televisivi e per dispositivi portatili. Ad esempio, mercati ritenuti promettenti sono quelli dei display di grandissime dimensioni, per luoghi pubblici, e dei display flessibili, adatti a nuovi tipi di applicazioni.

Un'ulteriore strategia per tentare di mantenere remunerativi i prezzi è differenziare maggiormente il prodotto: più che per le caratteristiche tecniche, il consumatore è attratto dalla sempre maggiore fedeltà di rappresentazione, e conseguente comunicazione di emozioni, e lo schermo può svolgere un ruolo determinante nel rendere "reali" gli eventi. Fattori di coinvolgimento sono senz'altro la dimensione e la quantità di dettagli con cui l'immagine è percepita dallo spettatore.

Si prevede che il numero di famiglie che assisteranno a programmi in alta definizione utilizzando apparati HDTV saranno 45 milioni, a livello mondiale, entro la fine del 2008.

2. HD-ready, HDTV ... 1080p

2.1 Informare il consumatore

In [1] si ricordava che il 19 gennaio del 2005 l'associazione dell'industria europea EICTA aveva annunciato i requisiti essenziali per etichettare gli apparati in grado di elaborare e visualizzare segnali ad alta definizione. Fra i requisiti essenziali *Hd-ready*: risoluzione nativa minima pari a 720 righe, formato d'immagine *widescreen* e la capacità di accettare come ingresso i formati 1280 x 720 pixel a 50 e 60 Hz progressivo (720p) e 1920 x 1080 pixel a 50 e 60 Hz interlacciato (1080i).

L'anno successivo, il 20 marzo 2006, ha introdotto il logo *HDTV,* che individua i dispositivi in grado di ricevere ed elaborare i segnali ad alta definizione diffusi per mezzo dei canali terrestri, via cavo e satellite o preregistrati, comprendendo quindi i STB (set-top-box) riproduttori e registratori DVD e televisori con ricevitore, al fine di visualizzarli direttamente o per mezzo di display HD-ready.

In questi anni le vendite di apparati HD ready sono cresciute costantemente: oggi sono più di 15 milioni in tutta Europa e più del 40% delle vendite è rappresentato da apparecchi di alta gamma (LCD, Plasma o a retroproiezione).

Sempre a partire dal 2006, alcuni produttori, con lo scopo di esaltare le caratteristiche del proprio apparato, hanno introdotto denominazioni quali *Full HD* e *1080*. In questi casi si è in presenza di slogan utilizzati per attrarre l'attenzione del consumatore sul fatto che i dispositivi sono in grado di visualizzare segnali a 1080 linee, senza che ciò implichi che questi siano a scansione progressiva, interlacciata, e neppure che la risoluzione orizzontale sia 1920 pixel, in alcuni casi limitata a soli 1440 pixel. Uno slogan, quindi, che non aiuta a certificare la definizione nativa del display, tanto meno la sua qualità, e che quindi può ingenerare confusione per il consumatore.

Pertanto il 30 agosto di quest'anno l'EICTA ha lanciato una nuova serie di loghi (figura 1) destinati a garantire che il prodotto etichettato

Fig. 1 - In alto il logo destinato ai display, compresi i televisori a plasma e LCD, in grado di ricevere, elaborare e visualizzare segnali ad alta definizione 1080p. In basso il logo destinato ad apparati capaci di di ricevere e decodificare segnali televisivi in alta definizione e dotati di un display di tipo "HD ready 1080p". (per maggiori dettagli: www.eicta.org)

con essi non solo abbia una risoluzione minima di 1920x1080 pixel, ma anche che le varianti a diversa frequenza di ripetizione di immagine [2] (24 Hz, 50 Hz e 60 Hz) possano essere acquisiti e visualizzati alla frequenza nativa, oppure superiore.

2.2 La scansione progressiva

Fin dalla nascita della televisione [3], un obiettivo fondamentale era l'aumento della definizione dell'immagine, in particolare incrementando il numero di righe, limitando nel contempo l'occupazione di banda del segnale trasmesso.

La scansione interlacciata, basata su due semiquadri, raggiungeva tale obiettivo, minimizzando il degradamento e garantendo una buona definizione verticale e risoluzione temporale. Gli schermi dotati di tubi a raggi catodici (CRT) per applicazioni televisive, con scansione interlacciata, hanno rappresentato per più di 50 anni, seppure con significativi miglioramenti tecnologici, una soluzione stabile e ottimale per la visualizzazione domestica.

Solo negli ultimi anni la produzione di CRT ha cessato di rappresentare la quasi totalità degli schermi e recentemente è stata quasi completamente abbandonata: oggigiorno, nei mercati più ricchi (Nord America, Europa e Asia), la domanda è quasi esclusivamente rivolta a schermi di tipo piatto (LCD, plasma, retroproiezione). Tali tecnologie sono strutturalmente adatte ad una scansione di tipo progressivo.

La scansione progressiva dell'immagine può offrire significativi vantaggi lungo tutta la catena di ripresa, codifica e trasmissione, ma, a parità di numero di righe, richiede di elaborare il doppio di pixel rispetto a quella interlacciata e quindi implica modifiche significative, e costi aggiuntivi, lungo tutta la catena. Le recenti tecniche di codifica digitale del segnale video, d'altro canto, presentano una maggiore efficienza nel caso di immagine progressiva rispetto a quella interlacciata e l'occupazione del canale, cioè il bit-rate necessario per la trasmissione o la memorizzazione, non raddoppia, proporzionalmente all'incremento in numero di pixel. La convergenza fra le tecnologie televisive e quelle informatiche, infine, favorisce l'adozione di un formato progressivo per le immagini.

Attualmente i programmi HD sono diffusi utilizzando il formato 1080i oppure 720p [2], ma l'industria si è impegnata negli ultimi anni per rendere disponibili apparati professionali di ripresa (telecamere), produzione e postproduzione in grado di operare in HD anche nel formato progressivo (1080p), e quindi con un bit-rate, a livello di interfacciamento di studio, pari a circa 3 Gb/s [4].

Analizziamo ora più nel dettaglio l'evoluzione degli schermi in grado di visualizzare immagini 1080p.

3. FDP (Flat Panel Display)

3.1 PDP (Plasma Display Panel)

I primi schermi al plasma (PDP) sono stati commercializzati a partire dal 1997 in Giappone e tuttora questa tecnologia raccoglie i favori del mercato giapponese, dove nel periodo aprile 2006 - gennaio 2007 la fornitura di PDP è cresciuta del 165% , percentuale superiore a quella

che ha caratterizzato i TV con schermo LCD, pari al 136%.

La competizione fra industrie giapponesi e coreane si evidenzia nella gara per conquistare le prime pagine grazie all'annuncio "presentato lo schermo più grande al mondo".

Infatti se era l'industria coreana ad annunciare alla fine del 2004 il pannello più grande (102"), è stata la giapponese Panasonic (Matsushita) a conquistare il primato nel gennaio 2006 con uno schermo da 103" (ovvero una diagonale di 2,6 m), pubblicizzato oggi anche dai grandi distributori specializzati italiani come "il display al plasma più grande del mondo", con ottime caratteristiche tecniche, un peso pari a 220 kg e un prezzo di 79000 €.

La realizzazione di pannelli così grandi implica difficoltà costruttive notevoli, sia dal punto di vista strutturale, che per il mantenimento su tutta l'ampia superficie di uguali caratteristiche di funzionamento (scarica stabile per tutte le celle) e di qualità dell'immagine. Questi prodotti, come anche quelli di dimensioni più contenute (da 70" a 80", già citati in [5]), sono evoluzioni della tecnologia alla base dei pannelli da 40" a 60", sviluppati per la visualizzazione di immagini a definizione standard e normalmente caratterizzati da formato d'immagine nativo pari a 1366x768 pixel. Questa tipologia di schermi 1080p non può ambire a soddisfare l'ampio mercato consumer a causa del costo.

I recenti miglioramenti tecnologici (si veda riquadro) hanno reso possibile la commercializzazione di plasma 1080p di dimensioni e costi che li rendono adatti al mercato consumer. Negli ultimi mesi del 2006, la Pioneer ha reso disponibile un PDP da 50" (a circa 8000 US$), grazie al dimezzamento delle dimensioni della cella (0,576 mm). Panasonic ha risposto, nel marzo 2007, con l'annuncio di PDP a 42". 50" e 58" (in aggiunta al già disponibile 65" e al già citato 103").

E' soprattutto la Panasonic a investire nella produzione di PDP: la sua capacità produttiva

PDP: Breve descrizione del funzionamento

Si basa sulla fluorescenza, emissione di luce da parte di fosfori. Strutturalmente, è costituito da una matrice di celle comprese fra due lastre di vetro. Ogni cella è costituita da tre sottocelle separate mediante costole (rib) perpendicolari allo schermo. Alle sottocelle corrispondono i fosfori rosso, verde e blu. Un campo elettrico è applicato ad un gas a bassa pressione contenuto nella sottocella: quando è applicata una tensione elevata si ha passaggio di corrente e il gas cambia stato, si ionizza e diventa plasma. Alcuni atomi del gas, eccitati, emettono raggi ultravioletti che colpiscono gli atomi di fosforo, questi ultimi, a loro volta, emettono energia sotto forma di luce visibile (rossa, verde e blu).

Le dimensioni delle sottocelle, rappresentano quindi l'ostacolo maggiore alla riduzione di dimensioni, consumi e costi per rendere il PDP a piena definizione un prodotto adatto al mercato consumer. Per ridurre tali dimensioni si agisce sullo spessore delle costole di separazione e sulle dimensioni di base della sottocella, compensata da un aumento della sua profondità, per assicurare comunque una ampia superficie coperta da materiali fluorescenti, e sull'impego di fosfori più efficienti: lo scopo è il mantenimento o l'incremento della luce emessa e la conseguente luminosità dell'immagine. I miglioramenti tecnologici hanno portato anche ad elevati valori di contrasto (fino a 4000:1) e di vita (occorrono più di 60000 ore di funzionamento, con immagini in movimento, prima che la luminosità si riduca a metà del valore iniziale).

Fig. 2 - Confronto fra la struttura waffle rib precedentemente adottata, a sinistra, e quella Deep Waffle Rib (fonte: www.pioneer.co.uk)

dovrebbe raggiungere le 960 000 unità al mese nel 2008, quando il suo quarto impianto produttivo sarà a pieno regime, mentre il successivo impianto, che sarà attivato nel 2009, dovrebbe farla crescere fino 1,96 milioni di pannelli al mese.

3.2 TFT-LCD (Thin Film Transistor - Liquid Crystal Display)

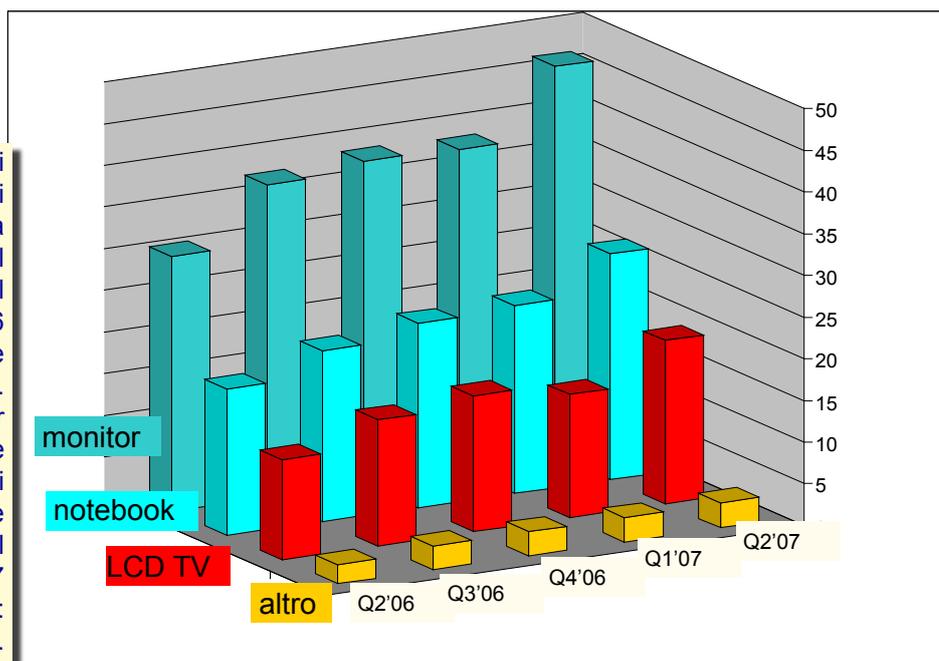
Anche in questo caso la competizione per il pannello più ampio è fra l'industria giapponese, (la Sharp commercializza dal novembre 2005 televisori da 65") e quella coreana (Samsung ha presentato nel 2006 un monitor da 70" in grado di riprodurre immagini a piena definizione con frequenza di quadro fino a 120 Hz e più recentemente ha presentato un monitor da 82", che dovrebbe essere disponibile sul mercato coreano nel prossimo novembre).

Questi ultimi schermi non sono ovviamente destinati, per dimensioni e per prezzo, al più ampio mercato consumer, a cui sono invece destinati i modelli 1080p di dimensioni più contenute (da 37" fino a 65").

Fig. 4 - il prototipo di "TV LCD più grande al mondo" dimostrato a Las Vegas a gennaio 2007 (fonte: www.sharp.co.jp)

La tecnologia LCD è predominante nelle vendite e nel 2006 ha rappresentato il 73% del totale degli incassi del mercato globale dei display; è in forte crescita per le applicazioni TV (figura 3),

In [1] si citava l'avvio della costruzione del secondo impianto della Sharp a Kameyama, il primo a utilizzare la tecnologia dell'ottava generazione. Proprio tale impianto ha prodotto il prototipo di LCD TV da 108" (figura 4), ovviamente "il più grande al mondo".

Fig. 3 - Milioni di schermi LCD-TFT di dimensioni superiori a 10" resi disponibili sul mercato mondiale dal secondo trimestre 2006 al secondo trimestre 2007.
Il numero di LCD per applicazioni TV è passato dai 12 milioni nel secondo trimestre 2006 ai 19,6 milioni del secondo trimestre 2007 (fonte: www.displaybank.com).

TFT-LCD:
Breve descrizione del funzionamento

LCD è una tecnologia di tipo trasmissivo, a valvola di luce. Mediante un campo elettrico viene fatto variare l'orientamento delle molecole delle sostanze liquide utilizzate, note con il nome generico di cristalli liquidi poiché hanno proprietà ottiche analoghe a quelle dei cristalli solidi. La polarizzazione della luce che transita varia seguendo l'orientamento delle molecole del liquido e quindi la luce passa, o non passa, attraverso due filtri polarizzanti disposti a 90° fra loro in funzione del campo elettrico applicato al cristallo liquido contenuto in ciascuna cella. Il dispositivo agisce quindi come una valvola di luce, e l'intensità luminosa può variare con continuità da un valore minimo, corrispondente al nero, ad un massimo corrispondente al bianco: si ottiene così la scala dei grigi. Il colore è ottenuto grazie al passaggio della luce attraverso un substrato di vetro su cui è depositato un filtro di colore corrispondente ai tre primari (rosso, verde e blu).

Nei display a matrice attiva (TFT) (figura 6) un film di semiconduttore è depositato su un sottile substrato di vetro e comanda ciascun pixel. I miglioramenti delle tecniche produttive, simili a quelle utilizzate per la produzione dei circuiti integrati, hanno consentito una riduzione nei costi, tempi e produttività: ad esempio si sono diminuiti, anno dopo anno, i passi di mascheratura, passando dagli 8 del 1995 ai 4 del 2000.

Fondamentale per le prestazioni e il costo dei display basati su TFT-LCD è la sorgente di luce. Mentre il tradizionale tubo fluorescente (CCFL, Cold Cathode Fluorescent Lamp) è tuttora oggetto di miglioramenti, vengono proposte possibili alternative quali nuove lampade (EEFL, External-Electrode Fluorescent Lamp, e FFL, Flat Fluorescent Lamp) o altre possibili sorgenti, LED (Light Emitting Diode) e laser.

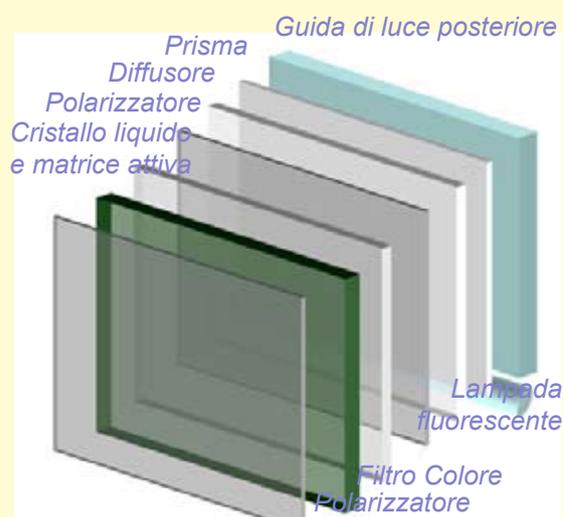



Fig. 6 - Struttura di un display TFT-LCD

E sempre la Sharp ha annunciato a fine luglio il progetto per la realizzazione del "Complesso di Produzione del 21° Secolo" per LCD a basso consumo di energia e di celle solari per la produzione di energia (il disegno concettuale del nuovo complesso è riprodotto nella prima pagina di questo articolo).

Nel marzo 2010, questo impianto sarà il primo a utilizzare la tecnologia del substrato di vetro della 10ª generazione. A partire da substrati di 2850 mm x 3050 mm sarà possibile realizzare fino a 8 pannelli della classe da 50" o 15 pannelli della classe 40".

Fig. 5 - Presentato in Giappone il 22 agosto il prototipo di "schermo LCD più sottile al mondo" (fonte: www.sharp.co.jp)

E proprio con tale tecnologia sarà possibile realizzare, nel 2010, "lo schermo più sottile al mondo" (figura 5) , caratterizzato da uno spessore di 20 mm (29 mm come spessore massimo), contrasto 100 000 : 1, peso 25 kg e consumo annuo pari a 140 kWh per i display nella classe 50".

Tra le caratteristiche con cui competere, oltre a grandezza e definizione, le industrie includono quindi anche quelle più importanti per l'ambiente, quali consumo e minor impiego di materiali, con un conseguente riduzione di ingombro e peso, fino a consentire di "appendere" al muro il grande schermo.

Fra i motivi che spingono i maggiori produttori che utilizzano la tecnologia LCD a enfatizzare la riduzione di spessore e consumi e l'incremento del contrasto, vi è senz'altro l'incombere sulla scena di una tecnologia molto competitiva, quella OLED.

3.3 OLED (Organic Light-Emitting Diode)

I produttori giapponesi Sony e Toshiba-Matsushita Display Technology (TMD) annunciano il loro impegno a sviluppare la tecnologica OLED per schermi di grandi dimensioni adatti ad applicazioni televisive.

La Sony ha dimostrato a gennaio uno schermo da 27" con la risoluzione 1080p, e ha annunciato, in aprile, l'intenzione di produrre entro l'anno uno schermo da 11" caratterizzato da una risoluzione pari a 1024x600 pixel. Le caratteristiche della tecnologia sviluppata dalla Sony (vedere riquadro nella pagina successiva) consentono di ottenere elevato contrasto (1000000:1), purezza dei colori ed elevata luminanza (superiore a 600 nit), tempi di risposta bassi, spessori ridottissimi (10 mm nel punto più sottile nel caso del pannello a 27", che si riducono a 3 mm per quello a 11").

La TMD, sempre ad aprile ha presentato in Giappone uno schermo da 20,8" (figura 7) basato sulla tecnologia LTPS OLED, da 20,8" con una risoluzione 1280x768 pixel. Questa tecnologia, di cui è stata pioniera la CDT (Cambridge Display Technology) utilizza polimeri che emettono luce (P-OLED) e ciascuno dei tre colori è depositato mediante un processo di stampa a getto d'inchiostro (*ink-jet*).

Toshiba prevede di avviare la produzione di pannelli per TV di classe 30" entro il 2009.

Anche l'industria coreana è attiva nello sviluppo della tecnologia OLED per TV e Samsung dimostrò un prototipo da 40" già nel maggio 2005.

Forse l'obiettivo più arduo per rendere commerciabile questa tecnologia è l'allungamento della vita: la durata dei materiali elettroluminescenti, ed in particolare quelli utilizzabili per il blu, non supera le 20000 ore. Una vita troppo breve rispetto a quella associata agli schermi al plasma e LCD. Si prevede che i rapidi progressi consentiranno, entro tre anni, di superare le 50000 ore, rendendo quindi competitiva la tecnologia OLED.

Quindi tre anni è ritenuto il tempo minimo per rendere competitiva, come prezzi e prestazioni, questa tecnologia. Occorre inoltre tener conto dei costi e tempi di realizzazione di nuovi impianti di produzione o di conversione degli attuali impianti da LCD a OLED. La DisplaySearch prevede che la quantità di pannelli OLED per TV potrà raggiungere i 3 milioni nel 2011, e crescerà a partire dal 2012, in concomitanza con la riduzione della richiesta di pannelli LCD.

Fig. 7 - Prototipo di schermo LTPS OLED da 20,8".

OLED:
Breve descrizione del funzionamento

Questi dispositivi si basano sulla emissione della luce da parte di materiali organici quando viene applicato un campo elettrico, cioè sul fenomeno della elettroluminescenza. La luce è emessa quando le lacune, cariche positivamente, provenienti dall'anodo e gli elettroni, carichi negativamente, provenienti dal catodo si combinano nello strato emissivo, composto di materiali organici.

Quindi la struttura convenzionale (figura 8) di una cella OLED, progettata al fine di massimizzare il processo di elettroluminescenza, consiste in una pila di strati: l'anodo, uno strato per il trasporto delle lacune, uno strato emissivo, uno strato per il trasporto degli elettroni, e il catodo.

La luce generata nello strato emissivo passa attraverso uno dei due elettrodi costituita da un materiale trasparente. Inoltre, utilizzando un materiale fortemente riflettente su una delle superfici dell'altro elettrodo, è possibile sfruttarlo come specchio, aumentando così la quantità di luce emessa.

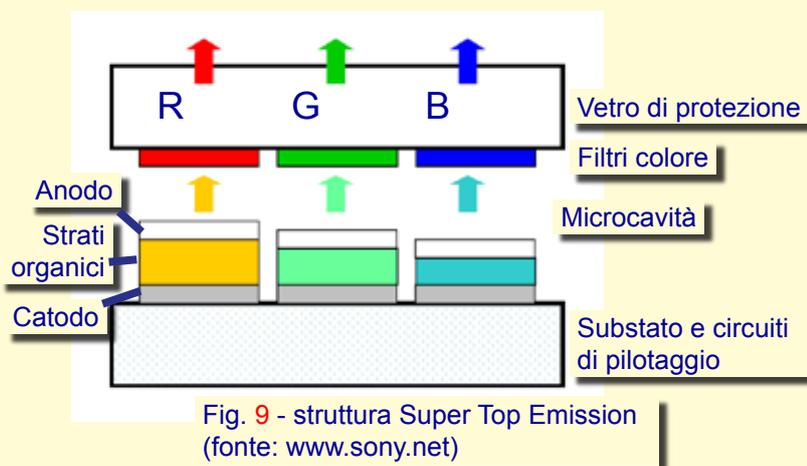
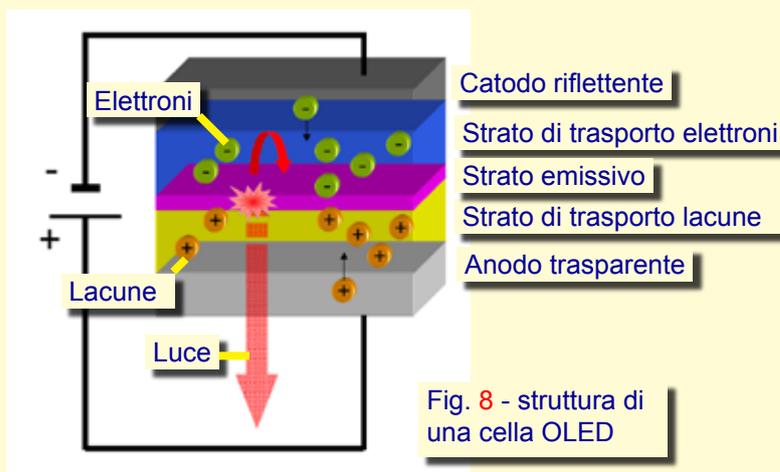
Se la luce è estratta passando attraverso il substrato (*bottom emission*), la luce passante è ridotta a causa della presenza dei circuiti di pilotaggio. Risulta quindi più efficiente la struttura che prevede l'estrazione attraverso il vetro di protezione (*top emission*), a contatto con l'anodo trasparente.

La struttura di base dell'OLED è molto semplice e quindi si può pensare che, a regime, i costi di produzione saranno fortemente competivi.

Le strutture effettivamente realizzate sono però più complesse, per garantire prestazioni elevate.

Ad esempio, la Sony ha sviluppato una struttura a microcavità (figura 9) che ha denominato *Super Top Emission.* Lo spessore dei film di materiale organico è ottimizzato per ciascuno dei tre colori; la cavità limita le interferenze garantendo una elevata purità; la combinazione dei filtri colore con la struttura a microcavità consente di ottenere un contrasto estremamente elevato.

Fig. 8 - struttura di una cella OLED

Fig. 9 - struttura Super Top Emission (fonte: www.sony.net)

4. Proiettori e RPTV (Retro Projection TV)

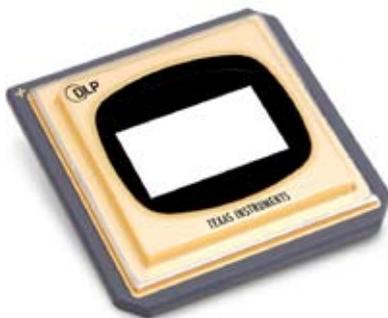
4.1 DMD™ (Digital Micromirror Device) e DLP® (Digital Light Processing)

Il primi prodotti basati sulla tecnologia DLP® furono commercializzati a partire dal 1996, a dicembre 2004 erano stati prodotti 5 milioni di apparati DLP, cifra che raggiunse i 10 milioni nel giugno 2006.

TV con risoluzione a 1080p (figura 10) basati sulla tecnologia DLP® erano stati annunciati nell'aprile 2005 [1] e oggi sono più di 70 i modelli di TV a retroproiezione con tale risoluzione elencati nel sito www.dlp.com, con dimensioni che variano da 50" a 73". Per quanto riguarda i proiettori, sono più di 50 quelli che presentano la risoluzione 1080p.

Uno dei limiti dei sistemi a retroproiezione è lo spessore: si può notare che sono commercializzati 11 TV nella categoria sottile a 1080p (Slim HDTV), a partire da un 56" (quindi una diagonale superiore a 1,4 m) e con una profondità inferiore ai 26 cm.

Recentemente l'accento è posto sulla realizzazione di TV a risoluzione 1080p adatti alla visualizzazione 3D [6].

Fig. 10 - Dispositivo 1080p (fonte: www.dlp.com).

4.2 TFT-LCD

La tecnologia LCD è utilizzata anche per proiettori e retroproiettori. In questo caso è necessario avere tre dispositivi, uno per ciascun primario, che operano come valvola di luce.

Incominciano ad apparire sul mercato anche sistemi con costi accettabili per usi domestici, a definizione 1080p (figura 11).

Fig. 11 - 26 marzo 2007: primo proiettore 1080p a scendere sotto il limite dei 3000 US$ (fonte: epson.mediaroom.com).

DMD: Breve descrizione del funzionamento

I sistemi di proiezione e retroproiezione realizzati con i dispositivi a microspecchi DMD™ e di preelaborazione dell'immagine realizzati dalla Texas Instruments sono denominati DLP®. La tecnologia DMD è basata sulla riflessione della luce da parte di una schiera di veloci specchi quadrati in alluminio (il diametro di un capello umano corrisponde a circa 4 microspecchi) ciascuno dei quali può assumere due posizioni in funzione della attrazione elettrostatica dello specchio e gli elettrodi connessi alla sottostante cella di memoria. La scala dei grigi viene ottenuta modulando la durata per cui ciascun specchio riflette o non riflette la luce. Nei sistemi con un solo DMD la luce monocromatica riflessa passa attraverso filtri di tre (i primari rosso, verde e blu) oppure sei colori (possono essere aggiunti anche i complementari, giallo, magenta e ciano) disposti su una ruota posta di fronte al dispositivo. Nei sistemi a tre DMD, caratterizzati da maggiori prestazioni in termine di luminosità ma più costosi, ogni dispositivo è associato ad uno dei tre colori primari.

5. La definizione cresce... anche sul palmo della mano

Anche se questo articolo è dedicato agli schermi adatti alla visualizzazione domestica di immagini 1080p, può essere interessante riprendere anche il tema affrontato due anni fa in [7] e valutare il progresso che ha caratterizzato i display per dispositivi *handheld*, quelli che, appunto stanno sul palmo della mano.

La tecnologia dominante per i display dei telefoni cellulari è tuttora quella LCD-TFT. Ed è ancora valida la considerazione che per i telefoni, con un peso intorno a 100 g e schermo di dimensioni limitate, ad esempio 2,4", la definizione adottata è quella pari a un quarto della VGA (QVGA, 320x240 pixel). Ma il successo commerciale di una nuova classe di telefoni intelligenti, dopo l'introduzione di iPhone (display 3,5", 420x320 pixel, 135g), dimostra l'interesse per dispositivi con schermi *thouchscreen* con dimensioni e peso comunque contenuti e capacità di collegamento ed elaborazione tali da trovare applicazione anche in ambiti fino ad ora esclusivi della televisione a definizione standard oppure dei notebook.

Fig. 13 - Prototipo presentato a maggio di OLED a colori su substrato flessibile: 2,5", 120x160 pixel, luminanza 100 nit, contrasto 1000:1, spessore 0,3 mm, peso 1,5 g (senza il driver) (fonte: www.sony.co.jp)

Nell'articolo citato [4] si accennava alla possibilità di realizzare sistemi a proiezione, per dotare dispositivi quali telefoni o PDA di immagini di sufficienti dimensioni ed ampiezza. Anche in questo ambito i prototipi recentemente presentati (figura 12) sembrano fornire soluzioni promettenti.

I progressi nel campo dei display flessibili sono notevoli, rendendo pensabile che in un futuro, anche se non prossimo, potremmo arrotolare il nostro schermo HDTV (figura 13).

Fig. 12 - Un esempio di microproiettore: la Microvision (www.microvision.com) ha annunciato alla fine di luglio un'accordo con la Motorola per sviluppare soluzioni basate sulla proiezione mediante laser. Il prototipo sarà basato sul dispositivo PicoP che è in grado di proiettare immagini WVGA, 854x480 pixel, in formato 16:9, a colori, integrabile in apparati di ridotte dimensioni (7 mm di spessore) e limitato consumo .

6. Definizione e qualità

I primi anni della storia della televisione videro una intensa competizione a livello mondiale per far crescere il numero di righe dei formati televisivi, ma, in pratica, la seconda metà del secolo scorso si è chiusa con una situazione abbastanza stabile: TV a definizione standard, quella ottimamente supportata dal tubo a raggi catodici, il display più economico e più diffuso.

In questi primi anni del terzo millennio, invece, la situazione sta evolvendo in modo estremamente rapido, grazie ai progressi tecnologici nella realizzazione di display con caratteristiche in termini di definizione, riproduzione colorimetrica, contrasto, peso, durata e soprattutto costo, difficilmente prevedibili, in base alle tendenze passate.

La presenza, a casa dello spettatore, di schermi televisivi con caratteristiche così differenti, ed in particolare in grado di visualizzare immagini con formati interlacciati o progressivi, a 1080i, 720p e ora 1080p, ha generato un vivace dibattito nell'ambito dei broadcaster [8] su quale siano i formati più adatti nell'ambito della produzione e in quello della diffusione.

La qualità finale dell'immagine HDTV non dipende solo dalla definizione nativa dello schermo su cui è visualizzata, ma tutta la catena di produzione e di codifica è determinante per garantire la fedeltà di riproduzione.

Proprio su questo tema, sarà pubblicato un articolo nel prossimo numero di Elettronica e Telecomunicazioni.

Bibliografia

1. M. Barbero, N. Shpuza: "verso l'Alta Definizione", Elettronica e Telecomunicazioni, anno 54, n. 1, aprile 2005
2. M. Barbero, N. Shpuza: "I formati HDTV (le raccomandazioni ITU-R BT.709 e BT.1543)", Elettronica e Telecomunicazioni, anno 54, n. 1, aprile 2005
3. M. Barbero, N. Shpuza: "Obiettivo 1000, Alta Definizione e schermi TV", anno 54, n. 2, agosto 2005
4. M. Barbero, N. Shpuza: "Intefacce Video", Elettronica e Telecomunicazioni, anno 55, n. 3, dicembre 2006
5. M. Barbero, N. Shpuza: "Display e proiettori, recenti progressi", Elettronica e Telecomunicazioni, anno 53, n. 2, agosto 2004
6. M. Muratori: "La tecnologia 3D-HDTV basata su DLP®", Elettronica e Telecomunicazioni, in questo numero
7. M. Barbero, N. Shpuza: "Grandi immagini sul palmo di una mano", Elettronica e Telecomunicazioni, anno 54, n. 1, aprile 2005
8. H. Hoffman: "HDTV - EBU format comparisons at IBC-2006", EBU Technical Review, October 2006

Che cosa è, come funziona:

La tecnologia 3D-HDTV basata su DLP®

ing. Mario **Muratori**

Rai
Centro Ricerche e
InnovazioneTecnologica
Torino

1. La tecnologia SmoothPicture della TI

La tecnologia SmoothPicture sviluppata dalla Texas Instruments (TI) ha lo scopo di produrre immagini "morbide" e simili a quelle ottenibili da pellicola cinematografica, utilizzando un Digital Micromirror Device (DMD) dell'ultima generazione (HD3) accoppiato con un attuatore ottico, ambedue inseriti in un visualizzatore a retroproiezione.

Questa tecnologia preserva l'originale piena risoluzione HDTV, in particolare il formato 1080p (progressivo), sugli assi principali (verticale e orizzontale).

Nelle prime due generazioni di DMD, i micro specchi di forma quadrata erano disposti secondo una griglia ortogonale, visualizzata in figura 1, e ognuno era dedicato alla proiezione di un singolo pixel sullo schermo di visualizzazione.

Sommario
La tecnologia DMD, basata sui micro-specchi, sviluppata dalla Texas Instruments (TI) per realizzare proiettori e retroproiettori a piena risoluzione HDTV (1080p) trova facilmente applicazione anche per la visualizzazione di materiale stereoscopico. In questa scheda è descritta la tecnologia SmoothPicture, sviluppata con il fine di ridurre il costo degli apparati 1080p. La multiplazione temporale alla base di questa tecnica, con l'aggiunta di un sincronizzatore per occhiali shutter, permette di realizzare un sistema di visualizzazione stereoscopica ad alta qualità (HD e assenza di flicker).

Per realizzare dispositivi capaci di trattare le risoluzioni maggiori, (p.es. l'HDTV 1920/1080p) contemporaneamente decrementandone il costo, TI ha sviluppato un DMD innovativo, nel quale i microspecchi sono disposti a losanga, cioè ruotati di 45° rispetto ai dispositivi delle prime generazioni, e sono posizionati secondo uno disposizione a quinconce[Nota 1] come illustrato in figura 2.

I nuovi DMD sono stati sviluppati per il formato 1280/720p e 1920/1080p. Quest'ultimo è composto da 540 coppie di righe di 960 coppie di colonne di microspecchi, come illustrato in figura 2.

Il numero di microspecchi risulta la metà rispetto ad una disposizione ortogonale di 1080 righe di 1920 microspecchi, ma offre una pari risoluzione orizzontale e verticale, sebbene la definizione in direzione diagonale risulti ridotta[Nota 2] .

Nota 1 - Nel gergo tecnico è nota anche come disposizione quincunx, utilizzando il termine inglese a sua volta di diretta derivazione dal latino. In linguaggio non specialistico si usa spesso la locuzione disposizione a scacchiera che però dovrebbe essere accompagnata dalla specifica del colore delle celle considerate: la scacchiera nel suo complesso ha infatti disposizione ortogonale.

Nota 2 - In una disposizione a quinconce di questo tipo ci si aspetta la presenza di una ripetizione spettrale centrata su 540 righe e 960 pixel, la quale potrebbe provocare dell'aliasing se la banda del segnale non venisse adeguatamente filtrata, riducendone di conseguenza la definizione diagonale. Tale filtraggio è infatti previsto nell'uso stereoscopico.

Nota 3 - Non tenendo conto dei pur necessari spazi tra i microspecchi, le dimensioni della matrice di microspecchi risultano solamente del 6% più ampie (circa il 12% in termini di area) nel DMD di nuova generazione HD3 rispetto a quelle del DMD della generazione precedente di risoluzione 1280/720p con disposizione ortogonale.

Inoltre, le dimensioni del chip del DMD 1920x1080 di terza generazione sono simili a quelle del chip per il formato 720p della seconda generazione (HD2)[Nota 3] . In questo modo, TI è riuscita ad offrire un dispositivo capace di gestire una risoluzione superiore ad un costo concorrenziale.

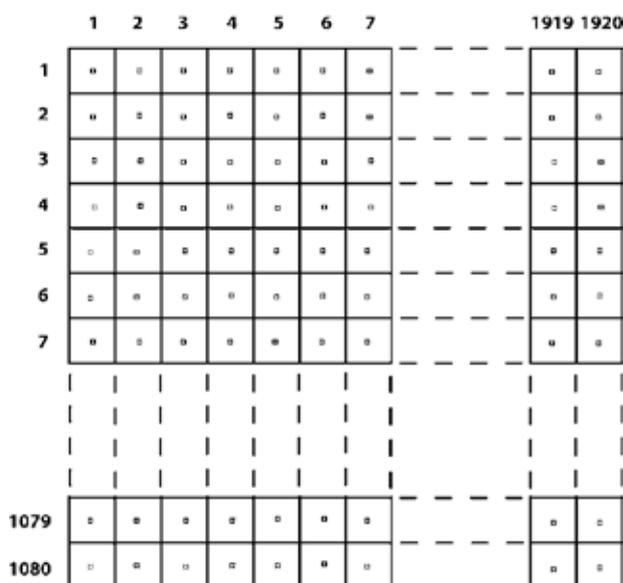



Fig. 1 – Disposizione dei microspecchi nei DMD delle prime generazioni (HD1, HD2, HD2+).

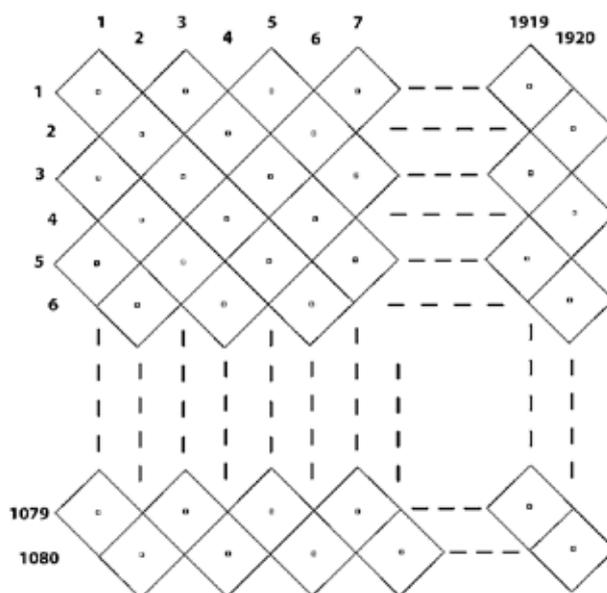



Fig. 2 – Disposizione dei microspecchi nei DMD della terza generazione (HD3).

Il DMD di terza generazione, avendo la metà dei microspecchi necessari, non potrebbe fornire da solo la piena definizione HDTV 1920/1080p.

Per questo motivo è coadiuvato da un attuatore ottico, sostanzialmente uno specchio basculante, che permette di orientare l'immagine prodotta dai microspecchi traslandola in orizzontale di ½ pixel, come illustrato in figura 3, così riproducendo sullo schermo del visualizzatore tutti i pixel costituenti l'immagine a piena definizione disposti secondo la consueta disposizione ortogonale[Nota 4].

La figura 4 illustra lo schema di funzionamento del sistema completo.

Il sistema ottico fin qui descritto necessita di una elaborazione, ancorchè minima, dell'immagine proiettata per poter funzionare correttamente. Infatti le immagini sono composte da pixel disposti secondo una disposizione ortogonale, mentre i microspecchi del DMD sono collocati secondo una disposizione a quinconce e sono in grado di proiettare solo la metà dei pixel costituenti l'immagine per volta.

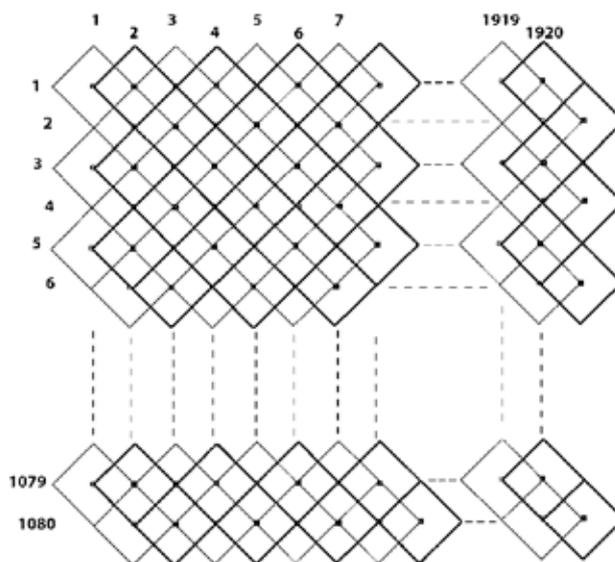
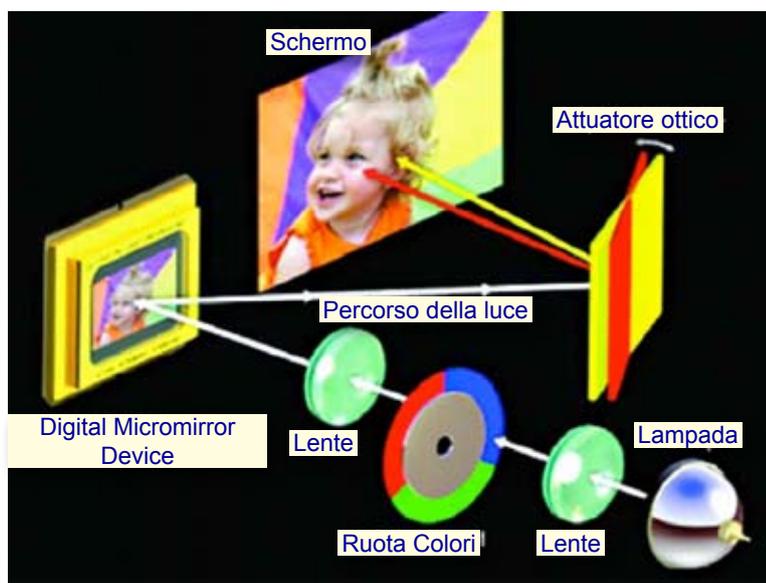



Fig. 3 – Risultato del basculamento di ½ pixel dell'attuattore ottico.

Fig. 4 – Percorso della luce in un apparato basato sulla tecnologia SmoothPicture (fonte: [3]).

Nota 4 - Ad essere pignoli, se le cose stessero veramente così - come peraltro riportato nella documentazione reperita [1, 2] - si avrebbe che il pixel più a sinistra delle linee pari (secondo la numerazione riportata in figura 2) non verrebbe mai illuminato. Perciò il lato sinistro dell'immagine totale risulterebbe seghettato; per ovviare a questo inconveniente potrebbe essere oscurato il pixel più a sinistra delle righe dispari, in corrispondenza alla posizione più a sinistra dell'attuatore ottico. In ogni caso la colonna più a sinistra risulterebbe incompleta o mancante.
Nello stesso tempo, il pixel più a destra delle righe pari nella posizione più a destra dell'attuatore ottico risulta in eccesso (formerebbe la colonna 1921) e probabilmente viene oscurato per rettificare il lato destro dell'immagine.
La perdita di una colonna di pixel in posizione laterale non è di importanza fondamentale e potrebbe essere evitata usando un DMD con una colonna in più; in ogni caso, dalla documentazione reperita su Internet non è dato conoscere qual è la situazione effettiva.

L'immagine da proiettare, che, si ricorda, consiste in un quadro televisivo in formato progressivo, viene scomposta in due semiquadri ricavati dall'immagine originale tramite sottocampionamento a quinconce secondo l'andamento illustrato in figura 5.

I semiquadri vengono passati al DMD in sequenza e vengono proiettati per una durata temporale pari alla metà del periodo di quadro (inverso della frequenza di quadro). In altre parole si effettua una multiplazione temporale dei semiquadri con relativo raddoppio della frequenza di presentazione dell'immagine (che diventa una frequenza di semiquadro, di valore pari al doppio della frequenza di quadro).

L'attuatore ottico bascula in sincronismo con il semiquadro proiettato di modo da proiettare i pixel nella posizione corretta e completare l'intero raster ortogonale dell'immagine a piena definizione nell'ambito di un ciclo, corrispondente ad un periodo di quadro.

Inoltre, la traslazione di mezzo pixel ammorbidisce i margini dei pixel, offrendo un'immagine più gradevole poiché diminuisce la visibilità dei bordi neri tra pixel dovuti agli spazi tra i microspecchi.

2. La tecnologia SmoothPicture e la stereoscopia

La tecnologia SmoothPicture effettua, in pratica, una multiplazione temporale di due semiquadri ricavati da un quadro in formato progressivo tramite un sottocampionamento a quinconce. Si noti che non si effettua alcun filtraggio per limitare la banda in direzione diagonale poiché l'immagine originale viene presentata nella sua completezza, e nella sua disposizione ortogonale originale, all'interno di un periodo di quadro.

A causa della presentazione di due sottoimmagini a frequenza doppia e dell'assenza di elaborazioni sull'immagine, tale sistema si presta a realizzare un apparato di visione stereoscopica basato sulla tecnica a multiplazione temporale con raddoppio della frequenza di quadro ([4]).

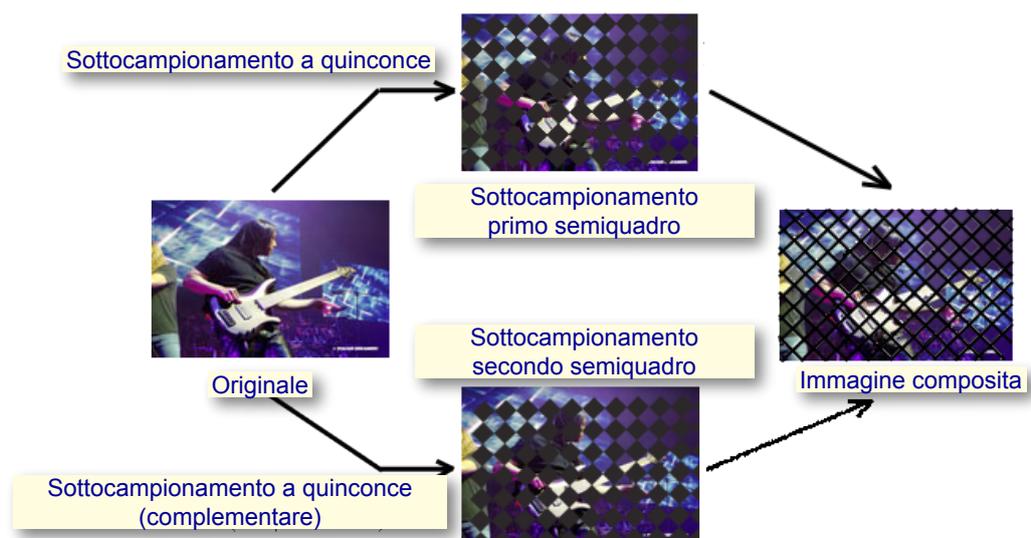



Fig. 5 – Schema di sottocampionamento dell'immagine proiettata.

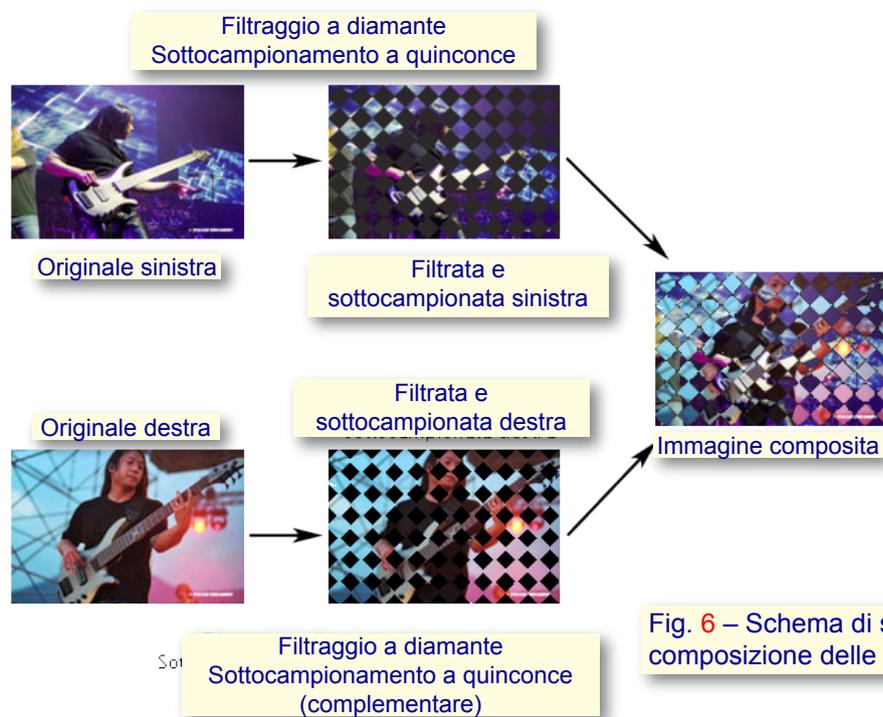



Fig. 6 – Schema di sottocampionamento e composizione delle componenti stereoscopiche.

In questa modalità, ogni immagine è ricavata dalla corrispondente componente la coppia stereoscopica tramite sottocampionamento a quinconce - con disposizione complementare - come illustrato in figura 6. In questo caso però è necessario un filtraggio a diamante per limitare la banda in direzione diagonale.

Le due immagini così sottocampionate sono composte in un'unica immagine in formato progressivo con la quale si alimenta un proiettore basato sulla tecnologia SmoothPicture (figura 6). Il visualizzatore effettua il processo descritto sopra, realizzando, di fatto, una proiezione stereoscopica in multiplazione temporale senza ulteriori interventi.

Il sistema viene completato da un dispositivo in grado di comandare degli occhiali attivi (occhiali shutter), per esempio un trasmettitore ad infrarossi operante secondo lo standard VESA[Nota 5], che permette di sincronizzare l'apertura e la chiusura degli otturatori sugli occhiali in sincronismo con la componente stereoscopica effettivamente visualizzata. Il segnale di sincronismo necessario è disponibile nell'apparato visualizzatore perché, in linea di principio, è lo stesso che controlla il basculamento dell'attuatore ottico.

Un ultimo accorgimento per l'utilizzazione del sistema è che a monte dell'apparato visualizzatore deve esistere un apparato che effettui il filtraggio e il sottocampionamento delle componenti stereoscopiche e la loro composizione in un'unica immagine in formato progressivo.

Si noti che il filtraggio a diamante necessario prima di un sottocampionamento a quinconce

Nota 5 - La Video Electronics Standards Association (VESA) è un ente di normativa internazionale fondato verso la fine degli anni 1980 dalla NEC Home Electronics e altri otto produttori di componenti ed apparati di visualizzazione. All'inizio lo scopo fu di normalizzare un display con risoluzione 800x600, in seguito l'interesse di VESA ha coperto altri settori. Per esempio, lo standard VESA-1997-11 definisce un connettore per la connessione di apparati di visualizzazione stereoscopici, quali shutter LCD, occhiali shutter, ecc. capace di fornire sia l'alimentazione, sia il segnale di sincronizzazione a tali apparati.

preserva la definizione lungo gli assi principali per l'occhio, ossia quello verticale e quello orizzontale, limitando la definizione lungo la direzione diagonale, dove però l'occhio presenta una sensibilità inferiore.

3. Realizzazioni pratiche della tecnologia DLP 3-D HDTV

Il sistema di visualizzazione viene commercializzato sotto la denominazione commerciale: DLP 3-D HDTV technology.

I dispositivi DMD sono prodotti solo dalla Texas Instruments, che li ha sviluppati.

Alcuni produttori di apparati di televisori a retroproiezione basati su DLP hanno inserito o stanno per inserire nei loro cataloghi alcuni modelli in grado di sfruttare la tecnologia DLP 3-D HDTV.

Alcuni produttori di accessori per la visualizzazione stereoscopica offrono soluzioni più o meno esplicitamente basate su elaboratore elettronico per la preelaborazione delle sequenze stereoscopiche e la visualizzazione con occhiali shutter[Nota 6]. Attualmente tale soluzione è l'unica disponibile sul mercato. Si noti inoltre che è prevista anche la funzionalità di generazione di sequenze stereoscopiche a partire da materiale video bidimensionale.

Dal punto di vista degli apparati video, sul mercato sono disponibili lettori in grado di supportare l'HDTV a piena risoluzione 1920/1080 in formato progressivo, ma a frequenze di quadro non superiori a 30 Hz[Nota 7], quindi inadeguati a supportare il formato di immagine necessario per la stereoscopia con la tecnologia DLP 3-D. Inoltre è lecito avanzare qualche dubbio sulla qualità ottenibile codificando immagini composite costruite come illustrato nel punto precedente.

4. Principali caratteristiche della tecnologia DLP 3-D HDTV

I vantaggi della tecnologia DLP 3-D HDTV possono essere così riassunti:

- ◊ La tecnologia DLP 3-D HDTV veicola ad ogni occhio un segnale a 60 Hz (complessivamente il sistema lavora a 120 Hz). Tale frequenza di quadro riduce il flicker che in alcuni sistemi (p.es. quelli basati sulla tecnica field-sequential) può essere particolarmente disturbante. Si noti che il sistema non risulta attualmente commercializzato in Europa, ma la tecnica del raddoppio delle frequenza di quadro sarebbe efficace anche nei sistemi europei a 50 Hz.
- ◊ La tecnologia ha un costo implementativo virtualmente nullo sui nuovi monitor DLP HDTV offrendo agli utenti un incremento futuro delle possibilità[Nota 8].

Nota 6 - Un'azienda commerciale presente su Internet specializzata in apparati per la stereoscopia vende bundle composti da occhiali shutter, trasmettitore IR per la loro sincornizzazione, pacchetto software (TriDef), ed eventualmente elaboratore elettronico, per la preelaborazione dell'immagine. Un'altra azienda invece propone un apparato apparentemente non dissimile da un normale lettore DVD, ma in realtà basato su elaboratore elettronico (Pentium Core 2 Duo, 2GB RAM, nvidia 8600GT), software dedicato (TriDef) per la conversione di sequenze 2D a 3D e la preelaborazione richiesta dalla tecnologia DLP 3-D.
Nota 7 - La risoluzione video massima per lo standard Blu-Ray Disc è l'HDTV 1920/1080/24p|50i|60i, mentre per lo standard HD HDTV è l'HDTV 1920/1080/24p|25p|30p|50i|60i.
Nota 8 - In [5] si evidenzia che i costruttori attualmente cercano di rendere più interessanti i loro prodotti offrendo la possibilità di espanderne in futuro la fruibilità. I prodotti basati su DLP 3-D potrebbero seguire questa tendenza, considerato che attualmente manca materiale video stereoscopico, a maggiore ragione specifico per il sistema, ed è piuttosto difficile, se non con costi aggiuntivi non tascurabili, godere di spettacoli stereoscopici.

◊ La tecnologia permette di sfruttare le funzionalità stereoscopiche del sistema tramite occhiali shutter, complessivamente ottenendo fedeltà colorimetrica ed elevata profondità di campo, nonché ottima reiezione al segnale indesiderato (pressoché totale assenza di effetto ghost).

5. Conclusioni

Una delle conclusioni più immediate è che l'industria dei display, quando riesce con poco costo, è attenta ad offrire apparati potenzialmente utilizzabili anche per la visione stereoscopica. La tecnologia DLP 3-D HDTV ne è un esempio: originariamente nata per realizzare display a piena definizione HDTV 1920/1080p con costi accettabili, viene proposta anche per la visualizzazione stereoscopica.

Ciò significa che non si ritiene ancora che la stereoscopia valga investimenti importanti, ma c'è un certo interesse per la visione tridimensionale, che si cerca di soddisfare appena è possibile.

Un secondo commento riguarda le possibili sorgenti di materiale stereoscopico. Anche dalle offerte commerciali cui si accenna in precedenza è chiaro che non esiste sul mercato del materiale video stereoscopico, a maggior ragione in formato HDTV. Si ricorre quindi alla ricostruzione tridimensionale a partire da materiale video bidimensionale[Nota 9] (giochi, film, video in genere).

A parte considerazioni legate alla produzione, di cui non si tratta se non per citare la volontà espressa da alcuni operatori del cinema di ricorrere sempre più alla produzione di materiale stereoscopico, si desidera qui far notare che i coraggiosi tentativi di proporre apparati di visualizzazione adatti alla stereoscopia che si sono recentemente registrati, si scontrano con l'assoluta indisponibilità di materiale video stereoscopico, anche perché non esiste alcun supporto adatto alla sua memorizzazione e alla sua fruizione (player).

Sembrerebbe pertanto ragionevole sostenere che uno dei fattori necessari alla diffusione della visualizzazione stereoscopica sia, oltre alla disponibilità di materiale video stereoscopico, la definizione di opportuni standard per la memorizzazione su supporto fisico, collegati ad opportuni standard che definiscano il formato video adatto e le funzionalità che un monitor deve avere per poter riprodurre materiale stereoscopico[Nota 10].

Bibliografia

1. David. C. Hutchison – The SmoothPicture algorthm. An overview – Digital TV Design Line, http://www.digitaltvdesignline.com/showArticle.jhtml?printableArticle=true&articleId=197007472
2. David. C. Hutchison – Introducing DLP 3-D TV – Texas Instruments, http://www.dlp.com/downloads/Introducing DLP 3D HDTV Whitepaper.pdf
3. DLP 3-D HDTV Technology - Texas Instruments, http://www.dlp.com/downloads/DLP 3D HDTV Technology.pdf
4. M. Muratori – Tecniche per la visione stereoscopica – Elettronica e Telecomunicazioni, n. 1, aprile 2007
5. M.Barbero, N. Shpuza – Alta definizione: display 1080p - Elettronica e Telecomunicazioni, in questo numero

Nota 9 - Con riferimento al software TriDef, ma anche all'apparato Virtual FX 3D converter commercializzato dalla i-O Display Systems, e alla blue box disponibile come accessorio agli schermi autostereoscopici Philips 3DWOW.
Nota 10 - Per esempio, molti engine presenti all'interno dei televisori a schermo piatto (plasma, TFT-LCD, DLP), al fine di migliorare la qualità dell'immagine, effettuano elaborazioni anche su più semiquadri vanificando la possibilità di visualizzazione con alcune tecniche, per esempio con tecnica field-sequential.

Che cosa è, come funziona:

Algoritmi e tecnologie per il riconoscimento vocale

Stato dell'arte e sviluppi futuri

Andrea **Falletto**

Rai
Centro Ricerche e
InnovazioneTecnologica
Torino

Sommario
Le tecnologie, sviluppatesi a partire dagli anni '50, per consentire il riconoscimento vocale si sono evolute nel corso degli anni e, grazie anche all'accresciuta capacità di elaborazione dei computer, trovano oggi sempre più ampia applicazione. In particolare possono costituire un valido ausilio anche in campi di stretto interesse degli enti televisivi, ad esempio per facilitare la sottotitolazione dei programmi, ed in prospettiva per consentirla anche in modo automatico e con basso tasso di errore. L'articolo è un'introduzione per meglio comprendere la complessità e le potenzialità di queste tecniche.

1. Agli inizi

Il Centro Ricerche e Innovazione Tecnologica è da sempre impegnato nello studio di nuove tecnologie in grado di fornire agli utenti disabili opportunità di inclusione e di fruizione dei programmi televisivi.

Le tecnologie basata sul riconoscimento vocale trovano impiego in questo ambito, per esempio facilitando la sottotitolazione automatica grazie al riconoscimento del parlato in un programma.

Questo articolo ha lo scopo di offrire una visione d'insieme dei processi coinvolti nel riconoscimento vocale tramite computer.

Anche se le maggiori innovazioni nel campo del riconoscimento vocale si sono sviluppate negli ultimi due decenni, la storia di questa tecnologia ha radici lontane.

Fig. 1 - Olivetti M20 – Commercializzato dalla Olivetti nel 1982 – Disponeva di un piccolo altoparlante per dare conferme acustiche e segnalare malfunzionamenti.

Fig. 2 - Un Notebook multimediale con Windows XP – I moderni software di riconoscimento vocale a dettatura continua – speaker dependent, possono funzionare in real time su una macchina come questa.

I primi tentativi di riconoscimento vocale vennero effettuati negli anni 50 negli Stati Uniti, con lo scopo di realizzare sistemi controllabili con la voce. I maggiori finanziatori della ricerca in questo campo furono la NSA (National Security Agency) e il dipartimento della difesa. Nel 1952 i Bell Laboratories svilupparono un sistema in grado di riconoscere i numeri da 0 a 9.

Il sistema poteva riconoscere solo parole singole: solo negli anni 70, presso la Carneige Mellor Univeristy venne messo a punto un prototipo con il quale era possibile riconoscere frasi complete, ma con un dizionario e una struttura grammaticale limitati. La potenza di calcolo richiesta per elaborare il riconoscimento era prodigiosa, prevedeva l'utilizzo contemporaneo di 50 computer.

Negli anni settanta i personal computer emettevano, tramite un piccolo altoparlante, alcuni suoni per dare all'operatore indicazioni sullo stato delle procedure e come feedback per indicare se la macchina stava operando correttamente. I PC non emettevano altri suoni o musica e solo alla fine degli anni ottanta comparvero le prime schede audio che permettevano la riproduzione di brevi suoni o frasi musicali (figura 1).

Negli anni 80 comparvero i primi dispositivi commerciali per il riconoscimento vocale. Nel 1982 la Covox commercializzò il Voice Master per Commodore 64 e successivamente per PC, in grado di realizzare una rudimentale sintesi vocale e un riconoscimento a parola singola in base ad un dizionario ristretto.

Nel 1982 Dragon Systems iniziò a produrre software per il riconoscimento vocale, seguita da IBM e Kurzweil. Da allora il mercato è stato invaso di applicativi, spesso venduti con un microfono a corredo, in grado di trasformare il proprio PC in una "stazione vocale".

Negli anni novanta i PC conquistarono la "multimedialità" che ha trasformato il personal computer in uno strumento versatile in grado di suonare, riprodurre CD e DVD ed elaborare segnali audio/video. Anche i notebook dagli anni

Fig. 3- Fase di addestramento (training) di un applicativo software di riconoscimento vocale: è richiesto all'utente di leggere un testo noto. In questo modo il software apprende le caratteristiche vocali e di pronuncia dell'utilizzatore.

novanta in poi diventano multimediali (figura 2), offrendo prestazioni analoghe ai modelli "da scrivania" (desktop).

Lo sviluppo degli algoritmi per realizzare applicazioni di riconoscimento vocale su PC ebbe una grande accelerazione proprio negli anni novanta. Oggi gli impieghi di questa tecnologia sono molteplici: si può comandare con la voce il proprio PC, il telefono cellulare, o il computer di bordo di un auto. Nel campo della telefonia, inoltre, i risponditori basati su riconoscitori vocali sono sempre più diffusi ed efficienti, si parla sempre più spesso di *call center* vocali automatici.

Il forte sviluppo della ricerca in questo campo ha permesso di realizzare software anche per il mercato consumer. Con questi programmi, trascorso un periodo di addestramento sulla voce dell'utente, si può dettare un testo parlando in modo naturale (riconoscimento vocale dipendente dal parlatore, *speaker dependent*, a dettatura continua). La precisione di riconoscimento di questi software è del 95 e 98 %. Il testo che state leggendo è stato scritto utilizzando un software di riconoscimento vocale.

2. Il riconoscimento vocale dipendente e indipendente dal parlatore

I sistemi di riconoscimento vocale, si dividono in due categorie: *speaker dependent* e *speaker independent*

- *Speaker dependent*: in questo caso il modello vocale viene adattato alla voce dell'utente. In pratica durante la fase di installazione, viene chiesto all'utente di leggere un testo con voce e velocità naturali (figura 3). Il sistema si adatta così alle caratteristiche della voce e della pronuncia dell'utilizzatore. Questi sistemi offrono i migliori risultati in termini di precisione permettendo anche, dopo un po' di pratica, di correggere gli errori di interpretazione tramite il microfono, usando solo le funzioni vocali. Gli algoritmi su cui sono basati, prevedono che venga tenuta traccia delle correzioni, per consentire di imparare dagli errori.
- *Speaker independent*: permettono il riconoscimento di un parlato generico, senza es-

sere legati ad un determinato timbro di voce. La precisione di questi sistemi è inferiore rispetto a quelli dipendenti dal parlatore. La loro applicazione principale si individua nei servizi di informazione automatici, in cui, ad esempio, tramite il telefono ci viene chiesto di dire il nome della città da cui intendiamo partire

Risponditore: *"Dica solo il nome della città da cui si desidera partire"*

Utente: *"Ancona"*

Risponditore: *"Lei ha detto"* - *"Ancona"*- *("?")*- *"dica si o no"*

Utente: *"si"*

Ogni individuo ha un proprio timbro vocale e un modo diverso di pronunciare le parole. I sistemi speaker independent offrono buoni risultati in situazioni in cui quello che viene detto dall'utente fa parte di una ristretta lista di parole oppure è prevedibile, come nel caso di risposte a scelta multipla.

3. Il Database - dizionario del software di riconoscimento

Il funzionamento di un sistema di riconoscimento vocale si basa sulla comparazione dell'audio in ingresso, opportunamente elaborato, con un database creato in fase di addestramento del sistema. In pratica l'applicativo software cerca di individuare la parola pronunciata dal parlatore, cercando nel database un suono simile e verificando a che parola corrisponde. Naturalmente è una operazione molto complessa, inoltre non viene fatta sulle parole intere ma sui fonemi che le compongono (figura 4).

I sistemi *speaker dependent* possono riconoscere correttamente oltre cento parole al minuto, confrontando quello che viene detto con un vocabolario di almeno 200.000 lemmi. Grazie al *training* sul parlatore, un normale PC è in grado di effettuare questa operazione in tempo reale, in background, e consentire all'utente di dettare un testo, estendendo le possibilità degli applicativi di acquisizione e trattamento testi.

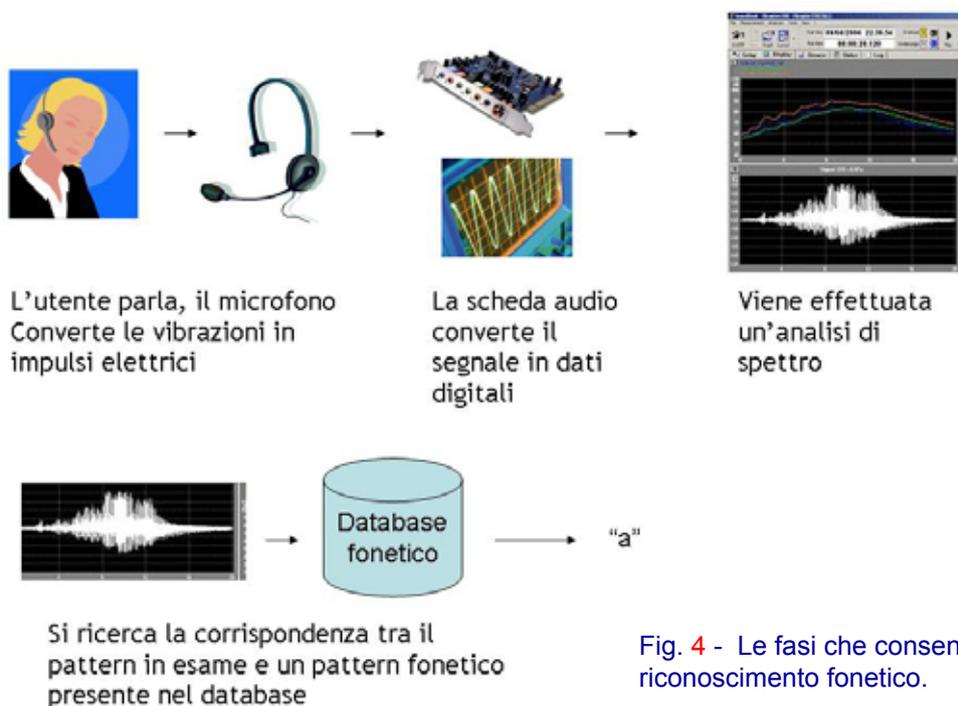



Fig. 4 - Le fasi che consentono il riconoscimento fonetico.

I sistemi *speaker independent*, come accennato, hanno una precisione inferiore perché non dispongono del modello vocale del parlatore. Per aumentare la precisione e operare su un database composto ad esempio da 200.000 lemmi, è necessario "insegnare" al sistema tutti i modi diversi in cui ogni singola parola può essere pronunciata. In pratica non potendo effettuare il *training* sul parlatore, la complessità di sposta verso il database che diventa molto grande e oneroso da costruire. Devono essere elaborate molte migliaia di ore di materiale audio con parole note, pronunciate da persone diverse.

Ovviamente il riconoscimento *speaker independent* richiede un computer estremamente potente oppure una elaborazione off line che può comportare, per una singola CPU, ore di elaborazione per riconoscere e trascrivere un minuto di audio.

4. Come funziona il riconoscimento vocale

La materia che si vuole trattare è molto complessa. In questa sede si è scelto di affrontare l'argomento con l'ottica di spiegare solo i concetti chiave, ricorrendo ove necessario all'utilizzo di esempi.

Il riconoscimento vocale automatico è basato su una sequenza di processi che si può così riassumere:

1. Trasformazione dei dati audio dal dominio del tempo al dominio della frequenza tramite FFT (Fast Fourier Transform)
2. Organizzazione dei dati ottenuti (tramite l'applicazione delle regole e del dizionario fonetico di una lingua)
3. Riconoscimento dei singoli fonemi
4. Composizione dei fonemi in parole e applicazione di un modello linguistico caratteristico della lingua in uso

Glossario

- Suono: Una vibrazione che si propaga in un mezzo (aria o solidi) come onda
- Fonema: l'unità minima che ha un valore distintivo all'interno di una lingua (a, t, e, z). Nella lingua italiana i fonemi sono circa trenta
- Parola: Un insieme di fonemi che uniti danno luogo a un vocabolo di senso compiuto in una certa lingua

Nel dominio del tempo il segnale audio si presenta come una forma d'onda periodica di frequenza compresa tra 300 e 3000 Hz (consideriamo la banda vocale). In questo formato le informazioni non permettono di individuare un pattern correlato a quello che viene detto. E' quindi necessario effettuare una conversione dei dati dal dominio del tempo al dominio della frequenza. Viene quindi realizzata una analisi di spettro del segnale, considerando una finestra di pochi campioni per volta e applicando la trasformata di Fourier. In questo modo è possibile identificare le frequenze che compongono il suono in esame e quale ampiezza ha ogni singola componente.

La FFT viene applicata tipicamente ad un segmento di audio della durata di un centesimo di secondo, dal quale si ricava un ipotetico grafico con l'ampiezza di ogni frequenza che compone il suono. Il riconoscitore vocale ha un database costituito da molte migliaia di questi "grafici" ognuno dei quali rappresenta l'enorme quantità di suoni diversi che la voce umana può produrre.

4.1 Dai grafici che rappresentano le frequenze componenti ai fonemi

Caso ideale

Procedendo per gradi, consideriamo ciò che avverrebbe in un caso ideale, immaginando cioè

che tutti abbiano la stessa voce e strumenti di analisi audio perfetti.

Il “grafico” del suono in analisi viene confrontato con tutto il database fino a che il sistema individua quello più simile. In questo modo è possibile stabilire che si trattava, ad esempio, di una “a”. In realtà il sistema dalla FFT ricava dei valori in base ai quali, per ogni centesimo di secondo, viene calcolato un *feature number*. Il *feature number* è quindi un numero che rappresenta il suono nel centesimo di secondo in esame. Anche il database contiene i grafici o *pattern* di riferimento sotto forma di numeri.

In una ipotesi ideale, ci sarebbe una corrispondenza diretta tra un *feature number* e un fonema. Quindi se il segmento di audio analizzato mostrasse come risultato il *feature number* n° 52 significherebbe che il parlatore ha pronunciato una “h”. Il *feature number* n° 53 corrisponderebbe ad una “f” e così via.

Sfortunatamente nel mondo reale le cose sono più complesse perché:

- Ogni volta che una persona dice una stessa parola, la dice in modo differente: quindi non produce mai lo stesso suono per ogni fonema. Per il nostro orecchio non costituisce un ostacolo, siamo infatti perfettamente in grado di capire un amico con il raffreddore, per il computer invece non è così immediato.
- I computer non dispongono dell’ascolto intenzionale (la caratteristica per cui il nostro sistema psicoacustico ci permette di “ascoltare” gli archi anche nel pieno d’orchestra) quindi i rumori di fondo, la musica e gli altri suoni sono un elemento fortemente disturbante e possono inficiare il riconoscimento del parlato in modo imprevedibile
- Il suono di ogni fonema cambia a seconda del fonema che lo precede e che lo segue. Il suono della “t” nella parola “tavolo” è molto diverso nella parola “antenna” o “treno”
- Il suono di un fonema cambia se si trova all’inizio o alla fine di una parola, ad esempio la “a” in “astice” è molto diversa dalla “a” in “roma”, quindi le due “a” producono sequenze di *feature number* molto differenti.

Tutto questo ci porta a descrivere il riconoscimento vocale nel caso reale.

Caso Reale

Iniziamo dicendo che un fonema dura molto di più di un centesimo di secondo. Evidentemente ogni fonema produce più *feature number*: se vene analizzato un centesimo di secondo per volta, significa che ogni secondo vengono prodotti 100 *feature number*. Se in un secondo viene detta la parola “mano”, dei cento *feature number* calcolati una parte rappresentano la “m”, una parte rappresentano la transizione dalla “m” alla “a”, una parte rappresentano la “a” poi la transizione dalla “a” alla “n” ecc.)

Visiti nel dettaglio, i passi per realizzare un riconoscitore sono.

- Istruire il software su come suona un fonema nelle varie pronunce e posizioni all’interno delle parole.
  - ◊ Un tool di training processa migliaia di registrazioni diverse dello stesso fonema.
  - ◊ Il sistema analizza ogni centesimo di secondo dell’audio e calcola un *feature number* in base all’ampiezza delle frequenze componenti.
  - ◊ Il sistema memorizza quindi migliaia di *feature number* per ogni fonema.
  - ◊ Maggiore è la quantità e l’eterogeneità del materiale analizzato, maggiore sarà la capacità del software di riconoscere le parole correttamente.
- Nel contempo, durante la fase di training, il software apprende anche una serie di dati statistici. Il dato più importante è costituito da quante probabilità ci sono che un determinato fonema generi una certa sequenza di

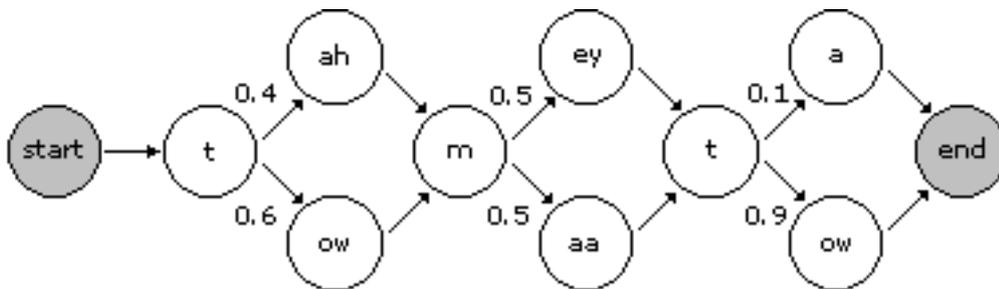



Fig. 5 - Esempio di Hidden Markov Model applicato alla pronuncia della parola inglese *Tomato*.
Questo modello prevede la pronuncia della parola in tre modi diversi
t ow m aa t ow - Inglese Britannico
t ah m ey t ow - Inglese Americano
t ah mey t a - Possibile pronuncia quando si parla rapidamente.
Si noti, dalle probabilità attribuite alle connessioni tra i fonemi, come il modello è leggermente sbilanciato verso il l'inglese britannico.

feature #. Prendiamo ad esempio il fonema "h". Il software dopo aver analizzato migliaia di "h"

- Apprende che in un centesimo di secondo del suono di una "h"
  - Ci sono il 55% di probabilità che compaia il feature #52
  - 30% di probabilità che compaia il feature #189
  - 15% di probabilità che compaia il feature #53
- Nel caso di una "f" ogni centesimo di secondo ci sono
  - 10% di probabilità che appaia un feature #52
  - 10% di probabilità che appaia il feature #189
  - 80% di probabilità che appaia il feature #53

Questa analisi delle probabilità è usata nella fase di riconoscimento; supponiamo che durante il riconoscimento i sei feature number calcolati durante sei centesimi di secondo siano:

…….52, 52 ,189 ,53 ,52 ,52…..

Il riconoscitore calcola le probabilità che sia uno dei trenta fonemi della lingua italiana.

Le probabilità che 52, 52 ,189 ,53 ,52 ,52 corrispondano ad una "h" sono

80% * 80% * 30% * 15% * 80% * 80% = 1.84 %

Le probabilità che il suono sia una "f" sono:

10% * 10% * 10% * 80% * 10% * 10 % = 0.0008 %

Il calcolo delle probabilità viene fatto per tutti i fonemi dalla "a" alla "z". Se tutti danno un risultato inferiore a 1.84%, molto probabilmente il suono analizzato è una "h".

Per mettere in pratica quanto esposto, i computer si affidano a strumenti matematici complessi. Tra i più usati a questo scopo c'è l' "Hidden Markov Model" HMM (figura 5).

In questo caso l'HMM viene usato per modellizzare una grossa matrice di fonemi, collegati tra di loro da "ponti" più o meno larghi, in base alle probabilità che un fonema sia correlato ad un altro.

I dati in input, come fossero un flusso di auto in marcia prendono preferenzialmente una strada o l'altra nella matrice a seconda di quanto i ponti tra un fonema e l'altro sono larghi e permettono il passaggio del traffico. La larghezza dei ponti viene modulata dalle statistiche calcolate dagli altri blocchi del sistema. Nel caso prima descritto il ponte che collegava il fonema precedente alla "h" era più grande (1,84) rispetto a al ponte che lo collegava alla "f" (0,0008).

4.2 La suddivisione dei fonemi

Il riconoscitore vocale ha anche bisogno di sapere quando un fonema finisce e inizia il successivo. A questo scopo vengono ancora impiegati gli "Hidden Markov Models".

E' possibile formarsi un idea di come funziona la procedura in questo esempio: supponiamo che il motore di riconoscimento individui una "h" seguita da un "ee". In base alle statistiche il fonema che corrisponde al suono "ee" ha:

- il 75% di possibilità di produrre un feature #82 per ogni centesimo di secondo,
- il 15% di possibilità di produrre un *feature* #98 e
- il 10% di dare origine a un *feature* #52.

Teniamo presente che il *feature* #53 appare anche nel fonema "h". Mettendo in fila i *feature number* di "hee" otteniamo:

……..23,52,52, 52, 189, 53, 52, 52, 82, 52, 82,82 ……..

. . . . h . . . . …….. . . . . . .ee. . . . . . .

Quindi dove finisce il suono "h" e inizia "ee" ?

Gli Hidden Markov Models

Sono modelli statistici che si possono applicare a sistemi che presentino la proprietà di Markov. Si dice che un sistema dispone della proprietà di Markov quando gli stati che può assumere in futuro dipendono solo dallo stato presente e sono indipendenti dagli stati passati.

Detto in un altro modo è un processo in cui le condizioni in un dato momento dipendono solo dalla situazione nello stato precedente e non da come si è giunti a tale stato. Come nel caso di un automobilista che decide di prendere una strada solo in base a quella che ha lasciato un momento prima, senza tener conto della mappa generale. Ciò che conta, alla fine è il punto in cui arriva e, visto a posteriori, il percorso che ha compiuto.

Hidden Markov Model hanno la caratteristica di poter determinare i parametri ignoti (hidden) in base a parametri osservabili.

Un esempio: un amico vive lontano e ama andare in moto. L'amico vi dice che un certo giorno

- andrà in moto verso la montagna,
- andrà in moto verso il mare
- oppure passerà la giornata al cinema.

La scelta di cosa farà è determinata esclusivamente dalle condizioni atmosferiche del giorno in esame. Non avete informazioni definitive sul meteo della città dove abita l'amico ma avete sentito per radio le previsioni per la sua regione. In base a cosa vi dirà che ha fatto, dovete indovinare com'era il tempo nella città del vostro amico nel giorno in questione.

Guardando i numeri appare che i 52 sono raggruppati all'inizio della sequenza gli 82 sono verso la fine. "Ad occhio" possiamo dire che la divisione si trova da qualche parte in mezzo ai due gruppi. Il computer per stabilirlo fa uso degli HMM: in questo caso ci sarà una matrice di *feature number*, i ponti avranno larghezze diverse in base alle probabilità (75% - 15% - 10%). In un primo istante il segnale di input percorrerà la strada che collega i feature number caratteristici della "h", poi i ponti saranno più agevoli nella zona della ee. Il segnale avrà effettuato dentro la matrice il percorso "hee". Le informazioni in uscita dalla matrice tengono traccia di tutto il processo e sarà possibile spere quanti millisecondi è durata la "h" e quanti millisecondi è durata la "ee")

4.3 Il silenzio

Il software sfrutta anche le pause del parlato per identificare l'inizio di un nuovo fonema, inoltre si occupa di analizzare in fonema "silenzio". In realtà le pause del parlato vengono sfruttate per prendere dei "campioni" del rumore di fondo e del fruscio che generano pattern di *feature number,* esattamente come il resto dell'audio.

Queste sequenze di dati vengono usate dal software per "depurare" l'informazione sonora utile dal pattern di rumore.

4.4 Uniamo i fonemi insieme, formando delle parole

Per l'algoritmo di riconoscimento tutti i processi descritti sopra rimangono nella sfera del possibile: tutto può essere rimesso in discussione e ri-calcolato fino a quando non si prova a mettere i fonemi insieme per formare le parole.

Il suono di un fonema cambia in relazione a quello che viene detto prima ed a quello che viene detto dopo. La "a" di "Aldo" "suona" diversamente da quella di "Andrea". Nella A di Aldo c'è una parte di "l" e nella A di "Andrea" c'è un po' di "n".

I software di riconoscimento risolvono questo problema creando dei tri-foni che sono terne di fonemi, composte tenendo conto del contesto. Quindi esisterà un tri-fono per il suono "silenzio-a-llll" e uno per il suono "silenzio-aaa-nnn". Considerando che devono essere contemplate tutte le combinazioni e che in italiano ci sono circa 30 fonemi, otteniamo 303 = 27.000 tri-foni.

Quelli simili vengono raggruppati allo scopo di ridurre i calcoli.

Durante i processi descritti il riconoscitore ipotizza più concatenazioni, basate sulle possibili combinazioni dei diversi fonemi. Calcola la probabilità che ciascun fonema sia nel posto giusto rispetto agli altri e stabilisce quale concatenzione ha maggiori probabilità di essere quella giusta. Ogni centesimo di secondo un nuovo *feature#* si aggiunge e integra le informazioni precedenti, fonema dopo fonema e silenzio dopo silenzio. Le combinazioni meno probabili vengono scartate, come pure le combinazioni impossibili di fonemi per la lingua in esame e viene scartata anche la possibilità che ogni centesimo di secondo inizi un nuovo fonema.

Riduzione dei calcoli e aumento della precisione

Dopo aver riconosciuto una parola dal punto di vista fonetico, sembra semplice trovare il temine corrispondente nel database. Può però accadere che: il parlatore non abbia pronunciato la parola chiaramente, che un rumore imprevisto abbia inficiato il contenuto dell'audio o che la divisione dei fonemi non sia avvenuta correttamente. Può succedere che "sono sotto casa anch'io" sia stato diviso male: "sonoso ttocasa anch'io".

Non trovando le parole corrispondenti nel dizionario, il sistema deve provare ad elaborare i fonemi in un altro modo ripetendo parti della procedura. Deve essere fissato un *time-out* o un numero di cicli, oltre il quale viene usata la parola più probabile, anche se sotto la soglia di sicurezza, oppure viene saltata la parola non riconosciuta per passare alla prossima. Alcuni

software indicano le parole non identificate con un simbolo, es. "…." o "???".

Per ridurre i tempi di calcolo e aumentare la precisione, vengono impostate delle regole per restringere il campo di ricerca. Si consideri che:

- Ci sono milioni di parole, ma normalmente ne vengono usate poche migliaia nella lingua corrente. La ricerca inizia da quelle più ricorrenti.
- Le regole grammaticali e linguistiche possono permettere di scartare combinazioni probabilmente sbagliate. Tra "eccomi sono qui" "ecco mi sono qui" invece di ricominciare confrontando entrambe le frasi con i feature # e i risultati degli HMM, viene, in base al modello linguistico, semplicemente scartata la seconda
- Con opportuni criteri, le sequenze di parole più comuni vengono memorizzate: la sequenza "Il presidente del consiglio dei ministri" è molto più probabile di "il residente del consiglio dei ministri"

Inoltre quando le parole da riconoscere non sono parte del lessico comune ma di un linguaggio specifico, deve essere caricato un dizionario appropriato. Per i software di riconoscimento consumer esistono dizionari con termini del linguaggio medico, giuridico, scientifico, tecnico ecc.

5. La modalità di composizione dei fonemi in parole in base ai contesti

5.1 Riconoscimento privo di grammatica

È il campo in cui i sistemi di riconoscimento vocale operano con maggiore efficienza.

Come descritto nell'introduzione, si tratta di contesti in cui il vocabolario e la struttura sintattica delle frasi da riconoscere sono limitati e in cui le scelte possibili sono previste a priori. Il software di riconoscimento può scartare tutte le risposte non contemplate. Anche nel caso di pronuncia non chiara la scelta che il software deve effettuare è semplice: se la parola pronunciata non è riconducibile alle parole che si aspetta di "sentire" la scarta e, se previsto, può chiedere di ripetere.

Un esempio:

Supponiamo di analizzare il caso di un sistema di domotica che disponga di un telecomando vocale.

L'utente conosce a priori quali comandi il suo sistema può accettare: ad esempio

Elenco comandi:

(Accendi le luci | telefona a |
invia una mail a | Componi)

Elenco parametri:

(in salotto | in cucina | nella scala |
Giovanni Marzio Patrizia |
0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9)

Sintesi vocale: "*Dire un comando"*

L'utente pronuncia un comando.

Il sistema, deve solo riconoscere i comandi sopra elencati e imporre regole prestabilite.

Le frasi che potrebbero scaturire da questo esempio sono del tipo:

Telefona a Patrizia

Invia una mail a Marzio

Accendi le luci nella scala

Componi 0621456690

Il riconoscitore è anche programmato per applicare un filtro sintattico: se l'utente pronuncia "*Componi*", il resto della frase non può che essere una sequenza di numeri. In questo caso il riconoscitore dopo aver identificato la parola "*Componi*" si aspetta di sentire dieci possibili vocaboli: *zero*, *uno*, *due*, *tre* … *nove*.

5.2 Dettatura discreta e Modello linguistico

Ormai la potenza dei moderni personal computer permette di utilizzare software di dettatura continua. Maggiore è la velocità del processore più rapidamente si può dettare, maggiore è la dimensione della memoria RAM disponibile, più alta è la precisione nel riconoscimento (i calcoli vengono fatti su porzioni di audio più grandi).

Per completezza si espone anche il funzionamento a dettatura discreta anche se non vi sono più molte applicazioni per questo tipo di tecnologia. L'argomento però introduce il concetto di modello linguistico, usato anche dai più moderni algoritmi per il riconoscimento della dettatura continua.

Nella dettatura discreta, il parlatore è privo di vincoli e può dire qualsiasi parola compresa in un dizionario che può essere grande a piacere (solitamente qualche migliaio di lemmi). Durante la dettatura deve essere lasciato uno spazio tra le parole per indicare al software quando una parola finisce e un'altra inizia.

Come già indicato, la sequenza di processi che permettono il riconoscimento automatico del parlato sono:

1. Trasformazione dei dati audio dal dominio del tempo al dominio della frequenza tramite FFT
2. Organizzazione dei dati tramite l'applicazione delle regole proprie del dizionario fonetico e della lingua in uso.
3. Riconoscimento dei singoli fonemi
4. Composizione dei fonemi in parole e in sequenze di parole mediante applicazione di modelli linguistici

Il modello linguistico

Gli algoritmi (*tool*) che compongono i database di riferimento e realizzano il training del software, oltre ad analizzare i fonemi, effettuano le analisi statistiche anche sulle parole, elaborando una grande quantità di frasi. Si parla di numeri dell'ordine delle decine di Giga Byte di testo, alcuni milioni di parole: per rendersi conto della mole di dati contenuti in questi data base, si consideri che tutta la produzione letteraria di Dante Alighieri occupa qualche dischetto da 1,44 MB.

Da questa analisi di milioni di frasi vengono ricavate delle statistiche. Per esempio, data una parola, viene ricercata l'evidenza di schemi che indichino se è frequentemente seguita da un'altra (es. "Il Santo" è spesso seguita da "Padre"). In pratica, in fase di utilizzo, quando il riconoscitore individuerà un qualsiasi vocabolo, mentre sta lavorando sui fonemi di quella successiva, ipotizza una lista di parole che potrebbero seguirlo, in base al calcolo delle probabilità. Quando ha finito il lavoro sui fonemi e individua la parola, la confronta con la lista. Se la parola è presente, la può considerare riconosciuta e procedere oltre. Naturalmente il processo funziona nei due sensi, sia per la parola che precede che per quella che segue. In realtà il processo non viene fatto sulla singola parola ma su terne di parole:

Realizzazione del modello linguistico

Prendiamo ad esempio questa frase:

> *"Sono le 8: gli operai metalmeccanici entrano nei cancelli della fabbrica."*

Viene convertita in

> *"Sono le otto duepunti gli operai metalmeccanici entrano nei cancelli della fabbrica punto"*

La frase viene suddivisa in terne di parole.

<inizio frase>
Sono le otto
le otto duepunti
duepunti gli operai
gli operai metalmeccanici

operai metalmeccanici entrano
metalmeccanici entrano nei
entrano nei cancelli
cancelli della fabbrica
della fabbrica punto
<fine frase>

Questo esempio, riguarda una frase sola per cui ogni terna di parole compare una sola volta. Su milioni di frasi invece, si delineano delle statistiche. Ad esempio la terna *"sono le otto"* o *"gli operai metalmeccanici"* compariranno più volte.

In un caso reale, guardando un segmento delle analisi vedremmo ad esempio statistiche di questo tipo:

….
….

Terna di parole	ricorrenze
città del Vaticano	3125
città di Roma	6122
città di toma	2
città di tufo	22

….
…..
altre terzine

La terna *"città di Roma"* compare 6122 volte, mentre *"città di tufo"* solo 22. Quando viene pronunciato "citta di…" ci sono maggiori probabilità che la parola successiva sia *"Roma"* invece di *"tufo"*. Se il pattern fonetico non è chiarissimo ma "assomiglia" a *"Roma"*, il software interpreta *"Roma"* e si occupa della parola successiva.

Naturalmente può accadere che una combinazione di parole non sia mai stata "sentita" da sistema e non sia nel suo database. Se l'utente pronuncia chiaramente e il riconoscimento fonetico termina con un risultato probabilistico alto, la frase potrebbe essere *"la città di scatole"*. In questo caso il riconoscimento "puro" ha prevalso sul modello linguistico. Naturalmente il software, aggiorna il suo database con questa nuova combinazione, indicando di averla sentita almeno una volta (o una volta in più).

L'utilizzo del un modello linguistico riduce notevolmente il tasso di errore, mediamente portando la precisione dall' 80% al 95%.

Utilizzo del modello linguistico, vantaggi e svantaggi:

- Il riconoscimento è più veloce, in quanto combinazioni improbabili vengono scartate a priori.
- Vengono risolte le ambiguità legate alle parole omofone, rare in italiano ma frequenti in altre lingue: ad esempio *"vizi"* e *"Vizzi"* o *"l'oro"* e *"loro"* possono essere pronunciate esattamente allo stesso modo. Come in inglese accade per *"to"*, *"too"* e *"two"*. In questi casi è solo il modello linguistico a determinare la scelta.
- Esistono anche alcuni svantaggi: se due parole appaiono simili, il software di riconoscimento sceglie la più probabile in base al training ricevuto. Purtroppo ci sono spiacevoli effetti collaterali, spesso imbarazzanti.

5.3 Dettatura continua

Il riconoscimento a dettatura continua permette di parlare con velocità normale, senza limitazioni di vocabolario. E' un processo molto più complesso perché il riconoscitore deve stabilire autonomamente l'inizio e la fine di ogni parola. Viene anche richiesta la pronuncia della punteggiatura.

La dettatura continua impiega strumenti matematici più complessi e, rispetto alla dettatura discreta, tutti i processi della struttura (FFT, analisi dei *feature#* per ogni fonema, divisione dei fonemi, combinazione in parole e modello linguistico) richiedono maggior complessità nei calcoli e una finestra di campioni più grande. La dettatura continua in tempo reale è possibile, in modalità *speaker dependent*, con un moderno personal computer. Per cifre dell'ordine del centinaio di euro, esistono in commercio software che offrono un riconoscimento molto preciso.

Invece, i software che lavorano in modalità a dettatura continua *speaker independent* presentano, ad oggi, forti limitazioni nella qualità del riconoscimento. La percentuale media è del 70 – 80 %.

Il testo ricavato da un riconoscitore *speaker independent* è utilizzabile per scopi statistici e di documentazione, in cui sia importante avere una traccia testuale del contenuto dell'audio ma non sia richiesta la precisione delle singole parole.

A esempio, attualmente non si può usare la dettatura continua *speaker independent* per ricavare dei sottotitoli automatici. Quando necessario, la soluzione adottata a tale scopo è di utilizzare personale addestrato ad ascoltare audio, elaborare in tempo reale una sintesi e pronunciare il sottotitolo. Un software *speaker dependent* opportunamente addestrato scrive il sottotitolo che risulterà corretto al 95-98%.

Ci sono forti interessi in gioco per quanto riguarda le tecnologie di riconoscimento vocale, sia da parte dell'industria, sia da parte del terziario sia da parte delle società produttrici di software per l'informatica consumer. Molti esperti sono convinti che il riconoscimento vocale costituirà il nocciolo dell'interfaccia uomo-macchina del futuro. Per questo motivo nuovi algoritmi vengono introdotti sul mercato frequentemente, tuttavia i processi impiegati sono riconducibili a quelli descritti in precedenza.

Una menzione particolare in questo campo meritano le reti neurali.

Le reti neurali

La limitazione dei computer è la loro rigidità. Anche se sembrerebbe assurdo parlare di rigidità in un sistema così flessibile, in realtà i computer non fanno altro che eseguire dei programmi. Un programma è definito come "una sequenza ordinata di istruzioni che eseguite sequenzialmente risolvono un problema". Il computer però non è in grado di adattarsi a situazioni nuove non

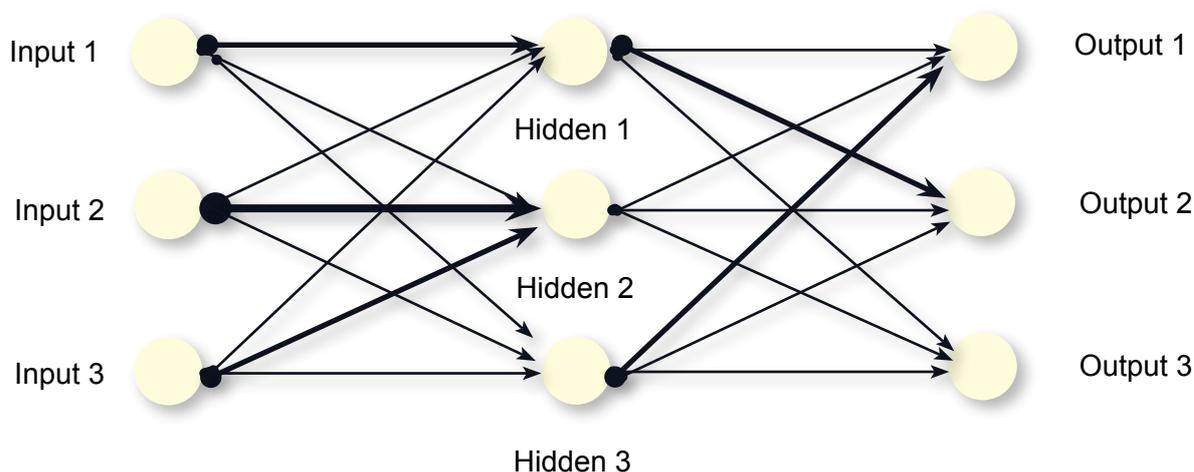



Fig. 6 - Esempio di rete neurale: la rete neurale si chiama in questo modo per similarità con le strutture che compongono il nostro cervello. I neuroni sono collegati da assoni e sinapsi. Nell'esempio il primo stadio è costituito dagli ingressi. I dati vengono convertiti in segnali compatibili con gli stadi successivi. Il secondo stadio, quello nascosto (hidden) elabora effettivamente i segnali. Il terzo stadio è quello di uscita e raccoglie i risultati adattandoli alle richieste del blocco successivo della rete neurale. Ogni collegamento e ogni stadio possono avere un peso ed una importanza maggiore degli altri. In questo modo la stessa rete può fornire risultati diversi con gli stessi input, esattamente come due persone reagiscono in modo differente in una situazione analoga. Il modo in cui vengono trattati i dati è caratterizzato dall'algoritmo che regola il comportamento della rete.

precedentemente codificate. Può rispondere ad uno stimolo se è stato programmato per farlo e anche le risposte sono previste a priori nel programma.

Nelle reti neurali (figura 6) ci sono molte unità di elaborazione indipendenti collegate tra loro. Il nome "rete neurale" è dovuto all'analogia con il nostro cervello, in cui i neuroni sono in grado di funzionare singolarmente e sono collegati tra loro tramite gli assoni e le sinapsi.

Nel nostro cervello gli impulsi elettrici viaggiano attraverso i collegamenti. Un neurone si attiva quando riceve un impulso sufficientemente forte. A sua volta il neurone emette un impulso elaborato in base a quello ricevuto inviandolo a tutti i neuroni ad esso collegati. I collegamenti tra i neuroni possono attenuare l'impulso modulandone l'intensità, fino a far si che in certe direzioni esso si spegna del tutto. Le reti neurali, analogamente, sono composte da tante unità collegate e da un algoritmo che può modificare i pesi (l'attenuazione) dei singoli collegamenti, in modo che il segnale di input prenda una certa direzione e porti ad un certo output.

Per addestrare la rete e migliorare l'algoritmo, si invia un impulso all'ingresso della rete e si osserva l'output. Si modificano poi i pesi dei collegamenti fino ad ottenere un output più vicino a quello desiderato. Si ripresenta un input, si valuta l'output e si ripete il processo finché è necessario. Una rete neurale, dopo la fase di addestramento, è in grado di fornire un output coerente, anche se riceve un input che non era stato presentato in fase di addestramento. Proprio per questo motivo le reti neurali trovano applicazione nel riconoscimento vocale e nel riconoscimento dei caratteri. In quest'ultimo caso si addestra la rete a riconoscere i caratteri finché la rete non distingue perfettamente tutte le lettere dell'alfabeto.

A questo punto dato un simbolo in input, la rete è in grado di stabilire a quale carattere somigli di più.

Conclusioni

Solo pochi anni fa i sistemi per il riconoscimento vocale presentavano un tasso di errore così elevato da renderli inutilizzabili nella pratica. Oggi invece i software di riconoscimento dipendenti dal parlatore offrono una precisione del 95 - 99% e sono in grado di "apprendere" dai propri errori. I software indipendenti dal parlatore, invece, sono mediamente meno precisi. Vengono impiegati per scopi statistici e di documentazione oppure nei call-center vocali automatici.

Un elemento critico, in grado di peggiorare fortemente il riconoscimento è la qualità dell'audio in esame o la presenza di rumori e/o musica.

Molti esperti sono convinti che il riconoscimento vocale costituirà il nocciolo dell'interfaccia uomo-macchina del futuro, per questo ci sono forti interessi in gioco che inducono il mondo dell'industria e della ricerca a perfezionare le tecnologie e gli algoritmi che permettano il riconoscimento a dettatura continua immune da errori.

Bibliografia

1. A tutorial on Hidden Markov Models and selected applications in speech recognition, L. Rabiner, 1989, Proc. IEEE 77(2):257--286.

2. What HMMs can do, Jeff Bilmes, U. Washington Tech Report, Feb 2002

3. Markovian Models for Sequential Data, Y. Bengio, Neural Computing Surveys 2, 129--162, 1999.

4. Acoustic Modelling - Microsoft Reseach www.http://research.microsoft.com/srg/acoustic-modeling.aspx

5. The most comprehensive site on Artificial Intelligence on the net http://www.generation5.org/